Exploring Novel Datasets and Methods for the Study of False Information

Edward Dearden, BSc (Hons)

School of Computing and Communications

Lancaster University

A thesis submitted for the degree of

Doctor of Philosophy

July, 2021

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. This thesis does not exceed the maximum permitted word length of 80,000 words including appendices and footnotes, but excluding the bibliography.

Edward Dearden

Abstract

Exploring Novel Datasets and Methods for the Study of False Information

Edward Dearden, BSc (Hons).

School of Computing and Communications, Lancaster University

A thesis submitted for the degree of *Doctor of Philosophy*. July, 2021

False information has increasingly become a subject of much discussion. Recently, disinformation has been linked to causing massive social harm, leading to the decline of democracy, and hindering global efforts in an international health crisis. In computing, and specifically Natural Language Processing (NLP), much effort has been put into tackling this problem. This has led to an increase of research in automated fact-checking and the language of disinformation. However, current research suffers from looking at a limited variety of sources. Much focus has, understandably, been given to platforms such as Twitter, Facebook and WhatsApp, as well as on traditional news articles online. Few works in NLP have looked at the specific communities where false information ferments. There has also been something of a topical constraint, with most examples of "Fake News" relating to current political issues.

This thesis contributes to this rapidly growing research area by looking wider for new sources of data, and developing methods to analyse them. Specifically, it introduces two new datasets to the field and performs analyses on both. The first of these, a corpus of April Fools hoaxes, is analysed with a feature-driven approach to examine the generalisability of different features in the classification of false information. This is the first corpus of April Fools news articles, and is publicly available for researchers. The second dataset, a corpus of online Flat Earth communities, is also the first of its kind. In addition to performing the first NLP analysis of the language of Flat Earth fora, an exploration is performed to look for the existence of sub-groups within these communities, as well as an analysis of language change. To support this analysis, language change methods are surveyed, and a new method for comparing the language change of groups over time is developed. The methods used, brought together from both NLP and Corpus Linguistics, provide new insight into the language of false information, and the way communities discuss it.

Publications

The following publications have been written during the course of this PhD. The contents of these papers constitutes part of the Thesis. In particular, the first two of these publications make up the majority of Chapter 3.

Edward Dearden and Alistair Baron. Fool's errand: Looking at april fools hoaxes as disinformation through the lens of deception and humour. In *20th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2019*, April 2019a. URL https://www.research.lancs.ac.uk/portal/en/publications/fools-errand(3fb53494-6b3a-4f21-9205-d525e87fa080).html

Edward Dearden and Alistair Baron. Fool's gold: Understanding the linguistic features of deception and humour through april fools' hoaxes. In *The 10th International Corpus Linguistics Conference, CL2019*, July 2019b. URL https://www.research.lancs.ac.uk/portal/en/publications/fools-gold(beda544a-cfa5-426d-9f82-f347b4ea4c50).html

Edward Dearden and Alistair Baron. Lancaster at SemEval-2018 task 3: Investigating ironic features in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 587–593, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1096. URL https://www.aclweb.org/anthology/S18-1096

Acknowledgements

As much as I would like to claim full credit for this work, that would be somewhat dishonest. It would also mean I would have nobody to drag down with me if one day I am exposed as a massive charlatan. So now I am going to try and thank everybody that I can think to thank. I will keep names to a minimum so I do not forget anybody, but you all know who you are. This has the added benefit that if I become mortal enemies with any of you one day, or you are ever revealed as a heinous criminal, I will not look like an absolute lemon for thanking you in my thesis.

Firstly, I would like to thank my supervisor, Alistair Baron, for being supportive throughout the PhD process, and for showing me how to do the research good. When I started my third year project under Alistair's tutelage, I could scarcely imagine the prospect of ever writing an incredibly boring and specific book – but here we are. It would not have been possible without his advice and general willingness to listen to me ramble on. I would also like to thank Paul Rayson for stepping in over the final 6 months and providing invaluable support and guidance. Gaining a writing-up PhD student when your plate is already rather full must be something of a challenge, so it is a testament to Paul's great generosity (or possibly insanity?) that he did it anyway. Thanks also go to Barry Porter for, along with Paul, always providing useful feedback at my progression panels, and reassuring me that I was not barking up completely the wrong tree.

Many other people from SCC have provided support throughout my PhD journey. Thanks go to the members of the NLP Group for making my job of organising the group remarkably easy, and for being a fine group of people with whom to discuss the fine art of NLP. Thank you also to the various characters of the PhD brew time brigade, who always provided a welcome distraction every Tuesday morning. Finally, thanks to the denizens of B55 for being the most reprehensible band of reprobates imaginable, and for making my time in Infolab a constant delight.

Last, but certainly not least, I would like to thank my friends and family, who have listened patiently to all of my complaining, and provided me with many welcome distractions from the world of PhD-ing. Thanks to all my fellow members of the Lancaster University Comedy Institute over the years, who made university life an absolute blast. And most of all, thank you to Hannah, for putting up with me over the entire process, and being an endless source of love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Deception is not a new concept. Lies have always been used as a tool to influence the opinions of others. However, with the ease of information propagation allowed by the internet, and social media, the proliferation of false information has become a major problem. With anybody able to start a news website, sites have appeared that provide news without getting caught up in ideas of journalistic integrity. To add to the problem, social media allows rumours to be created and shared at the click of a button. This is only made worse by the business model of the web, which rewards the attracting of clicks and views with little concern about the quality of information or the consequences of lies. False information, both intentional ("disinformation") and unintentional ("misinformation") spreads like wildfire in this environment.

In 2016, the idea of "fake news" exploded in popular discussion. False information was pointed to as a contributing factor to the outcomes of two major democratic events that year – the UK's EU referendum, and the 2016 US presidential election [Rose, 2017]. Suddenly, everybody was talking about how dishonest journalists and Russian propagandists were poisoning the well of knowledge, and spoiling democracy for everybody. Organised campaigns of disinformation were exposed [ODNI, 2017], and social media companies initially dragged their heels in doing anything about it[1]. Journalists from organisations such as `politifact.com` and `snopes.com` did their best to fact-check claims made by key figures, but the volume was simply too

[1] `https://www.bbc.co.uk/news/technology-37983571`

much for human fact-checkers to handle alone. The Covid-19 pandemic in 2020 further exposed the scale of false information, and the damage it caused, with widespread 5G conspiracies [Bruns et al., 2020], and vaccine scepticism [Johnson et al., 2020].

As mentioned, false information and "fake news" are not new [Soll, 2016]. Fake stories had dire consequences dating back to the middle ages with the blood libel – anti-Semitic conspiracies that resulted in the deaths of countless Jewish people. Whenever a new communication medium appears, it is inevitably misused to spread false information and propaganda. For example, when modern newspapers emerged in the 19th century, low quality tabloid journalism became rife, with fake stories being spread to gain advantage over the competition[2]. This was gradually dealt with through the introduction of journalistic norms, which aimed for balanced reporting. The internet has allowed the proliferation of false information on an unprecedented scale, which has outpaced these norms. However, society is yet to adapt to it, and we are lacking in media literacy, and solid policy, to mitigate the problems it causes. Long term solutions, such as education, will undoubtedly prove to be the most effective solutions[3].

In the meantime, various technical solutions have been proposed to help us stem the tide of false information. Automated fact-checking is one of the most popular, the idea being to programmatically verify claims made online [Thorne and Vlachos, 2018], either completely automatically, or as a support system for journalists. Other approaches have tried to identify false information based on language use [Rashkin et al., 2017], which theoretically allows an instant response even in the absence of facts. Identifying the users on social media who spread false information, is another area of work [Guo et al., 2020]. These techniques all attempt to help tackle the problem, in conjunction with more human interventions such as education and policy changes.

Over the course of this thesis, *false information* will be used as an umbrella term that refers to information that is factually inaccurate. *Disinformation* is defined as information that is intentionally deceptive or misleading, and *misinformation* as false information that is not intended to deceive. Sometimes, people have used *misinformation* as an overall term, with *disinformation* as a subset, but we will avoid this because it leads to ambiguity. *False information* will be used when encompassing

[2] `https://history.state.gov/milestones/1866-1898/yellow-journalism`

[3] `https://www.theguardian.com/world/2020/jan/28/fact-from-fiction-finlands-new-lessons-in-combating-fake-news`

both varieties, and mis/disinformation when being more specific. Finally we mention *fake news*, a term that has been appropriated by politicians to discredit journalists that criticise them, and has hence become somewhat meaningless. Despite this, we will occasionally use this term to refer to misleading news articles, as *fake news* is still the most succinct and universally understood way to refer to them.

1.2 Problem Statement

Some inherent problems exist with the idea of "detecting" false information. Firstly, it is very difficult. If it was easy to tell if one was being deceived, it would not be such a problem. This means that automatic systems are unlikely to get good enough to be confident that they are always right. There is also a potential for censorship. As soon as a "fake news detector" exists and is relied upon, what would stop somebody else from making one which simply censors information they disagree with? Similarly, would it mean that individuals or news publishers who share false information would be forever flagged as "fake"? These problems highlight the need for explainable systems. Without understanding the reasons behind a decision, predictions on the veracity of information is, at best, flawed and, at worst, useless.

There are also problems with the datasets used in current research. Firstly, there are too few standard datasets, making the comparison of results difficult. The range of information sources is also limited. Due to ease of collection, popular social media sites such as Twitter are often targeted. While these are important places to study, they only represent a narrow sample of the world's population. Another problem is constrained topics. Much work in news verification uses sources such as Snopes and Politifact, meaning that most of the articles relate to politics, and hot-topic issues such as climate change and vaccination. This limited variety of topics calls into question the generalisability of our understanding of false information.

Though there is a body of work dedicated to researching conspiracy theories [e.g. Uscinski, 2018, Douglas et al., 2019] and false information [e.g. Guo et al., 2020, Oshikawa et al., 2020], much of this work has focused on mainstream social media platforms, as opposed to niche communities dedicated to the topics[4]. While websites

[4]Though there are various exceptions, described in section 2.4

such as Twitter and Facebook may be where "normal" people are exposed to false information, it is not where the dedicated communities of like-minded individuals meet to discuss their ideas. By looking at forums, and other online communities, dedicated to the discussion of such topics, we can better understand the way they operate and the arguments put forward in spreading rumours or conspiracy theories.

Another limitation of current work in false information is that language is usually treated as static. The way the language of false information changes over time has not received much focus. This is a problem in much of NLP, and there has been a move to address it in recent years. By looking at language change in disinformation, we can better understand the way that deceptive tactics change, and adapt to new trends or events. Particularly, we are interested in how the make up of these communities changes over time, and whether the language of users changes accordingly. Developing techniques that deal with language change in false information communities may help to observe potential phenomena such as the migration of trolls into a community, or the indoctrination of new members to a conspiratorial belief.

The final problem in current research that we will discuss is that the distinction between misinformation and disinformation is widely neglected. Many works focus on the veracity of false information, not accounting for the intention of the author writing it. There are many reasons why an author may sincerely or insincerely spread a lie or a rumour. It is reasonable to think that somebody who sincerely believes that NASA faked the moon landings would communicate their ideas differently from somebody who is knowingly lying. Therefore, when performing linguistic analyses, it is naive to treat all false information as the same regardless of why it has been written. This is a difficult problem to solve, but it needs to be highlighted as a problem and kept in mind in the building of false information datasets.

1.3 Research Questions

RQ1: How can we increase our understanding of the language of false information by looking at previously unstudied sources? False information spreads across a range of genres and media, so it is important that we understand how generalisable the linguistic features of it are. This will involve looking at datasets of false

information which have not been studied in previous NLP research. We will also look at different types of linguistic feature, relating to style, topic, and deception. We will look for similarities between different forms of false information to see how generalisable the features are. In addition, we wish to use simple and explainable methods where possible, so we can gain insight into the language of false information, rather than black box techniques. For this we will use methods from Natural Language Processing as well as Corpus Linguistics.

RQ2: What methods allow us to observe the language of groups within communities, particularly regarding language change over short time-spans?
To understand the language of false information, we need techniques to explore the communities that discuss it. This will increase our understanding of the way they communicate and the sub-groups and individuals that make them up. Examples of sub-groups are groups based on ideological positions, or status within a community. Primarily, we are interested in language change, and the relative change of sub-groups to each other over time. Looking at this may highlight the influence different groups have on each other over time. In the context of false information, we can use these methods to better understand the communities where disinformation ferments. We will adapt existing methods for exploring sub-groups and individuals within the context of the wider community. These methods will be applicable to a range of communities, be they on or offline.

RQ3: What features characterise false information communities, and how do they compare to other communities? In answering this question, we hope to explore the makeup of users within such communities. We will study the language usage within them, observing how they discuss their theories. By comparing these communities to related communities, as well as communities not relating to false information, we will be able to learn more about where they sit within the broader online space. This will also feed into RQ1, by comparing the language of communities built around false information to those that are not.

1.4 Thesis Structure

This thesis consists of six main chapters, corresponding to a literature review and five substantial bodies of work. **Chapter 2** will survey literature relating to the topics relevant to this thesis. This will cover a range of areas, but will predominantly focus on Natural Language Processing (NLP), Corpus Linguistics, the study of online communities, and language change. A background in these topics will provide context for understanding the studies presented in the following chapters.

Chapter 3 will describe the creation of a corpus of April Fools hoax news articles, and its subsequent analysis. This will provide an interesting case study of false information, where the texts are verifiably false, and the authors do not believe what they are writing. We will perform a thorough analysis of the features that best characterise these articles, using methods from NLP and corpus linguistics. The hoaxes will also be compared to fake news, to see how generalisable their features are. This will contribute to answering RQ1, by seeing which linguistic features are useful for classifying two different types of false information.

In **Chapter 4**, we will produce a toolbox of methods for looking at language change, using UK parliamentary debate as a case study on which to test them. Initially, we will discuss a selection of methods that may be useful for looking at sub-communities (political parties) within larger communities (UK House of Commons). We will apply these methods to the parliamentary dataset to see what we can learn about the changing language of MPs over the course of Brexit debates from 2015 to 2020. This chapter will help to address RQ2 by testing several methods for comparing groups within a community over time. The toolbox will also equip us to examine false information communities, which will help contribute to RQ3.

Chapter 5 will propose a novel method, based on cross-entropy, for analysing the language of sub-groups over time. This method will build on the techniques described in Chapter 4, and will be designed specifically to compare the relative change of sub-groups, compared to each other, over time. An analysis of Brexit discourse, using the Hansard dataset from Chapter 4, will be performed to test the method and highlight its benefits and limitations. The proposed technique will contribute to RQ2.

Chapter 6 will introduce the first large scale dataset looking at the Flat Earth community, consisting of a Flat Earth forum, as well as a number of Flat Earth

subreddits. A thorough meta analysis will be completed of these groups, contributing to RQ2. This will be followed by a linguistic analysis, with the aim of learning more about the language of Flat Earth debate and answering RQ1 and RQ3.

In **Chapter 7**, we will apply language change methods from Chapter 4 to look at how the language of the forum from Chapter 6 changes over time. This will involve a comparison of the community to other related communities, as well as non-flat-earth communities. We will also identify groups within the forum based on posting behaviour and linguistic features, which will contribute towards answering RQ2.

Finally, **Chapter 8** will summarise and conclude the thesis. Here, we will describe the main contributions of the thesis, and evaluate how well the RQs have been addressed. We will also highlight areas for future work.

Chapter 2

Literature Review

This thesis will cover a range of topics over its course, and this chapter serves to provide a basis in each of these areas. The methods used in this work are mainly taken from the areas of Natural Language Processing (NLP) and Corpus Linguistics (CL), which will be described in Sections 2.1 and 2.2. While the focus will be on NLP and CL, the subject of this thesis is deeply cross-disciplinary, and we will touch on work from other areas such as social network analysis, psychology, sociology, forensic linguistics, and digital humanities. Section 2.3 discusses language change; another topic we will cover while answering RQ2. Adjacent to this work is a large body of existing work in online communities (Section 2.4), which will be important for answering RQ3. Finally, in Section 2.5, we will cover false information research, which brings everything mentioned so far together.

2.1 Natural Language Processing

Natural language processing (NLP) is a field at the intersection of linguistics and computer science, which involves the study of language using computers. Another term often used in relation to this is *computational linguistics*, which is sometimes considered as a slightly separate field [Clark et al., 2013], with NLP more focused on the engineering side of things compared to computational linguistics. In this thesis we will treat these concepts as interchangeable, mainly using the term NLP. Goldberg [2017] defined NLP as follows:

> "Natural language processing (NLP) is the field of designing methods and

> algorithms that take as input or produce as output unstructured, natural language data."

This is the definition that we will work with throughout this thesis.

Dealing with natural language is challenging because language is complex, ambiguous and messy. Though linguistic rules guide the construction of language, there is still much ambiguity, and the relationship between words is not self-evident from the words alone. That is not to mention the huge variety of languages that exist, each with their own rules and quirks.

NLP systems aim to overcome these challenges, and solve real world problems using natural language data, which is readily available – especially since the dawn of the world wide web. As well as solving these problems, NLP also includes much work focused on creating methods to model or process language as well as possible.

If we were to describe a pipeline of NLP, which most tasks slot into, it would be as follows:

1. **Data gathering** – This can involve challenges such as reading text from images, or scraping information from the web.

2. **Preprocessing and Normalisation** – This involves taking messy, unstructured data as input and turning it into something structured and usable. It may also include annotation.

3. **Feature Extraction** – Turning processed text into features that can be analysed or handed to a machine learning algorithm. This step can be handled manually or automatically.

4. **Analysis** – Making predictions, or reaching conclusions, on the basis of the text features.

While these stages are debatable and leak into each other, the general idea is that NLP tasks can take place anywhere between where text is found, and the high level analysis eventually performed on it.

One cannot talk about NLP without also discussing machine learning, which has become intrinsically woven into the fabric of NLP. This is because it allows the creation of models using sparse and complex data. In recent years, the field has been dominated

by deep learning methods, which can learn incredibly complex behaviours from large datasets, and have outperformed the previous state-of-the-art for many NLP tasks.

NLP can be used for many tasks. Firstly, there are the problems that concern the technical aspects of NLP systems – e.g. creating new methods to parse or annotate text. Machine translation is another example of a popular task. Aspects of a text can be predicted using NLP methods, such as sentiment or stance. As well as predicting things about language, NLP systems are also used for generating language, for example when making chatbots. NLP also has inter-disciplinary applications, as it can be used to answer research questions in the social sciences, humanities, and health.

This section aims to provide a basis in the regions of the NLP landscape relevant to this work. For more general information about NLP, there are several comprehensive text books [e.g. Jurafsky and Martin, 2009, Goldberg, 2017].

2.1.1 Preprocessing and Normalisation

Natural language can be non-standard, and difficult to process for all sorts of reasons. For example, there may be mistakes caused by image-to-text conversion, variation in spelling, or incorrect grammar. The increasing use of web-text also introduces a whole new range of factors to address: e.g. more text from non-native speakers, emojis, and URLs within the text. This means that an important step in NLP is to preprocess non-standard text into clean and usable data. This process is often overlooked, but issues at this stage can dramatically affect results further down the pipeline.

Tokenisation

Tokenisers segment text into *tokens*. Typically, a token is a word, but sometimes you might want to treat a multi-word expression (e.g. “New York”) as a single token. The simplest way to tokenise, in a language such as English, would be to identify every string delimited by white space. In reality, this is a naive system which does not work in many common cases (e.g. punctuation). More success can be enjoyed using regular expressions. The NLTK Package [Bird et al., 2009] includes a tokeniser which uses a deterministic function based on regular expressions. Such methods are quick, but have been improved upon in more recent NLP packages [Qi et al., 2020b, Honnibal et al., 2020]

Tokenisation in languages that do not use white space to delimit words, such as Chinese or Japanese, can be more challenging. Methods to tokenise such languages do exist [Asahara and Matsumoto, 2000], though in some cases, for Chinese at least, it has been suggested that working at a character level is more effective [Li et al., 2019b]. This goes to show how much these types of methods are dependent on language. In much NLP work, there is overly a focus on certain languages, particularly English.

Normalisation

As well as splitting text into words, there are certain steps that one might take to normalise one's data. Two common normalisation steps taken in NLP are stemming [Porter, 1980] and lemmatisation [Müller et al., 2015]. Both of these methods reduce words to their root form, stemming does this using only the characters of the provided word (e.g. "booking" to "book"), while lemmatisation takes into account the syntax, and sometimes also the intended meaning of the word (e.g. "better" to "good"). Another common normalisation step is removing punctuation, and other non-alphabetic characters such as emojis. Strings such as URLs or hashtags may be converted to tags.

These methods can be very useful, as they make the text data more dense, which can make it more usable. This only applies in certain tasks, however. For example, if one was predicting the topic of a text, lemmatising words makes a lot of sense. However, if one was trying to identify the author of a given document, this process might remove features of an author's style – whether somebody says "good" or "best" may provide clues about their personality. Normalisation steps should be chosen in a way that is appropriate for the task of interest, and there is no universal silver bullet implementation.

Annotation

Sometimes you may want to annotate your text with additional information. A common type of annotation is Part-of-Speech (PoS) tagging, which assigns grammatical tags (e.g. "ADJ", "NOUN") to words. Many PoS taggers have been developed. Early taggers used Hidden Markov Models [Garside, 1987], while more recent ones have used neural network based methods [Akbik et al., 2018]. Though the methods used have become much more complicated, the performance of these systems has not increased all that substantially [Manning, 2011], plateauing at around 97%.

Other annotation techniques can be used to capture different information. For example, the semantic tagger USAS [Rayson, 2008] assigns semantic labels to tokens, which can be used to help analyse the concepts discussed within a text. Named Entity Recognition (NER), meanwhile, is used to identify entities (e.g. people, places) within a text [Yamada et al., 2020, Wang et al., 2019]. An example of a use of NER would be seeing which public figures are discussed within a certain community. Various NLP Packages contain implementations of NER [Honnibal et al., 2020, Qi et al., 2020b].

2.1.2 Representations

An important consideration for any natural language processing task is the representation to use for texts. Traditionally, the main approach was to create sparse vectors based on counts of certain words or features in a text. The most basic example of this kind of model is the Bag-of-Words (BoW) representation [Harris, 1954]. In BoW models, documents are represented by sparse vectors which represent feature counts. Often these features are lexical features such as word counts, but they are also commonly word or character ngrams[1]. Complex information such as word order and context are discounted. Despite this simplicity, for many tasks BoW is still useful and can produce good results.

An extension of BoW that still enjoys reasonable success is tf-idf [Salton and McGill, 1986], a statistic which aims to measure the importance of a word rather than simply its raw frequency. The advantage of using tf-idf is that it removes the impact of common but uninteresting words and boosts the significance of rarer, more informative words. This is very useful in tasks such as topic modelling. However, when looking at the style of texts, less content-related features such as function words can be very informative [Pennebaker, 2013]. More detail on the choice of features for investigating language style can be found in Section 2.1.5.

Another common representation of documents is to choose a set of features via manual feature engineering. This involves choosing features using knowledge of the data. For example, when looking at style one may decide that pronouns [Pennebaker, 2013] or character trigrams [Peng et al., 2003] are useful features. Document level features such as type/token ratio and text length may also be employed. An advantage

[1]Ngrams are sequences of n consecutive items. For example, “Hello” could be split into character trigrams, “Hel”, “ell”, “llo”. The same can be done for words in a sentence.

is that the feature sets are very small and dense compared to more data-driven methods (e.g. BoW) which means models using them can run very quickly. Feature Engineering can also allow for greater interpretability of models, as the features are based on intuition and therefore if a feature performs well, it is sometimes easier to understand why. The disadvantage of feature engineering is that you ideally require some expert knowledge of the application domain and some would argue that humans may not pick up on some of the deeper features that a machine might identify. Feature Engineering has fallen out of favour slightly with the increasing adoption of deep learning approaches that learn their own feature representations [Tenney et al., 2019]. However, it can still be a very useful technique, which avoids being as much of a black-box as other methods.

Count based methods have a problem of sparsity. This can be addressed by using dimensionality reduction techniques such as Principal Component Analysis (PCA) [Abdi and Williams, 2010] and Latent Semantic Indexing (LSI) [Deerwester et al., 1990]. These methods reduce the total number of features, and also ensure that certain aspects of feature relations are maintained. Reducing the dimensions can also be useful in clustering, as these algorithms often struggle with high dimensional data [Kriegel et al., 2009]. Reducing the number of dimensions down to two or three also makes it possible to plot data in a straight-forward manner. Another advantage of dimensionality reduction is that it can help avoid overfitting. The main disadvantage is that it can make results more abstract and difficult to interpret, as one has to work out which features each dimension represents.

More recently, neural representations have become popular. Popular systems such as Word2Vec [Mikolov et al., 2013] and BERT [Devlin et al., 2019b] use neural networks to produce word representations. Word Embeddings provide dense vectors to represent words and generalise much better than count-based representations [Bengio et al., 2003, Mikolov et al., 2013]. The general idea is to train a model to transform words into a fixed size vector space which places similar words near to each other. This model creates vectors in such a way that similar words have similar vectors. Words are seen as similar if they appear in similar contexts, according to the distributional theory of semantics [Harris, 1954, Firth, 1957]. The most popular implementation of word embeddings is that of Word2Vec [Mikolov et al., 2013], though others exist such as GLOVE [Pennington et al., 2014]. Word embeddings have been shown to perform well

for classifying short texts due to their generalisability but for longer texts tf-idf still performs well [Shahmirzadi et al., 2019].

In the past couple of years, contextual embeddings have been widely adopted and outperform normal word embeddings. Unlike word embeddings, contextual embeddings produce vectors based on each individual word occurrence's context. This means that words with multiple meanings would get multiple different vectors. In normal word embeddings, e.g. Word2Vec [Mikolov et al., 2013], each word in the vocabulary corresponds only to a single vector. ELMo [Peters et al., 2018] was the implementation of contextual embeddings that gained popularity, though it has since been eclipsed by BERT [Devlin et al., 2019b].

A popular extension of word embeddings is document embeddings [Le and Mikolov, 2014, Dai et al., 2015]. These are vectors that are created to represent an entire document. They are trained alongside word embeddings, meaning that each document vector has some knowledge of all the words within a document. Document embeddings capture features such as topic of a text and can be meaningfully compared. Some works have used Document Embeddings for stylometry tasks [Agun and Yilmazel, 2017, Markov et al., 2017] where style is more important than topic, suggesting that document embeddings can capture elements of style as well as content.

Finally, it is worth considering traditional topic modelling techniques such as Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. LDA is a technique which generates a given number of topics for provided text. Each topic is a probability distribution over the vocabulary. This means that each word is represented by a vector where each value is the probability the word is in a certain topic. One can assign a topic to a document by looking at the topic of the words within said document. By looking at the words that characterise each topic, we can learn about what each topic represents. The advantage of LDA is that it is unsupervised, so one does not need to provide topic labels in advance, merely the number of topics. LDA can also be used as a form of dimensionality reduction, as each document can be represented by a probability distribution over all topics, rather than all words as would be the case in BoW.

There are, however, some significant shortcomings of LDA. The technique suffers from probabilistic issues leading to results which are not robust across multiple iterations [Chuang et al., 2015]. LDA is also not very explainable [Gillings, 2016].

One can produce a list of words with the highest probability of being in a certain topic, but these words do not always make sense and can be hard to interpret. Finally, and most significantly, LDA does not work well on short texts, e.g. social media posts. Yan et al. [2013] tried to alleviate this problem by finding topics from the corpus at large rather than using individual documents.

2.1.3 Machine Learning

Machine Learning is so widely used in NLP that it is crucial to have a basic grounding in current methods. This section provides a brief overview of machine learning used in NLP.

In Machine Learning, models are created that learn and infer characteristics and patterns within some data. There are two main divisions that ML methods fall into: Supervised and Unsupervised. We will not discuss the other two categories, Reinforcement [Sutton and Barto, 2018] and Semi-supervised learning [Zhu and Goldberg, 2009], as they are not relevant to this work.

Supervised Machine Learning

In supervised Machine Learning, the model is trained with labelled data and then evaluated on unseen, unlabelled data. The most commonly used technique of this nature is classification, where a model is trained to distinguish between multiple classes by learning relationships between the features of each inputted document and their labels. An example of this sort of task in Authorship Attribution would be training a model to choose which of two authors a text is written by [Jockers and Witten, 2010]. The model would be trained on extracts of text for which the author is known and then made to make predictions for texts that have no assigned author.

Linear models, such as logistic regression, are commonly used for classification. These models are fairly simple but only capture linear relationships between features and therefore do not pick up on more complex, non-linear patterns. Despite this, they are more understandable and computationally cheaper than many more complex methods. More powerful linear models also exist, such as Support Vector Machines [Boser et al., 1992]. In the case of SVMs, they get around the linear relationship limitation by finding linear relationships in a much higher dimensional space, but at the cost of additional

complexity and less interpretable results. SVMs still enjoy state-of-the-art performance in certain tasks.

Another common type of ML method are tree based models, such as decision trees and random forest, which are simple but struggle with sparse data. Some algorithms such as K-Nearest Neighbour have also been used which use similarity to learn the different classes.

More recent NLP work has made use of non-linear models, such as Neural Networks, as these can capture deeper, non-linear features. One reason for their popularity is that they learn their own feature representations and do not need to be given hand-crafted features. This means fewer assumptions need to be made in advance about the data. Neural Networks may be powerful, but they lack in explainability, as we discuss in Section 2.1.3. In NLP a particular variety of Neural Network, the Recurrent Neural Network (RNN) [Elman, 1990] has become popular. This is because of its ability to make predictions based on the neighbouring sequence of words. LSTMs [Sundermeyer et al., 2012] are a particularly successful variety of RNN which remember relevant words from the past and forget less relevant ones. Transformers [Vaswani et al., 2017] are another neural network architecture which have enjoyed success in NLP in recent years [Devlin et al., 2019b]. Though powerful, these complex methods require a significant amount of computation, and concern has been raised over problems such as raising the bar to entry in terms of computing resources, the environmental costs of such large models [Strubell et al., 2019], as well as various ethical issues linked to model size [Bender et al., 2021].

Unsupervised Machine Learning

Unlike supervised learning, unsupervised learning does not take labelled data as input. These techniques usually require a large amount of structured data to be successful. One of the main unsupervised machine learning tasks is clustering, where the model assigns each item into groups that it has created.

There are several popular methods for clustering. K-means clustering is a simple and very wide-spread method which clusters items by creating K means and allocating each item to its nearest mean, according to some distance metric. Spectral clustering [Ng et al., 2001] is a method in which dimensionality reduction is applied to the eigenvectors

of the similarity matrix before clustering is applied. Both these methods require the number of clusters to be specified in advance. Agglomerative Hierarchical Clustering [Nielsen, 2016] is another commonly used method where items are clustered based on distances between each other. For this method, instead of specifying the number of clusters, one picks a similarity threshold to determine the number of clusters. The mean-shift clustering algorithm [Fukunaga and Hostetler, 1975, Comaniciu and Meer, 2002] requires no number of clusters be provided. Each technique has different pros and cons, and may work better for certain "shapes" of data, or numbers of clusters.

There are various ways to assess the quality of clusters [Baarsch and Celebi, 2012], though none are foolproof. The Calinski-Harabasz index [Caliński and Harabasz, 1974] is one measure that can be used to assess the quality of clusters, with a higher score suggesting better defined clusters. This method, however, requires ground-truth labels, which are often not available. In cases where there are no labels, measures such as the Davies Bouldin index [Davies and Bouldin, 1979], or silhouette coefficient [Rousseeuw, 1987] are useful. These metrics can also be used to predict the number of clusters for a given set of documents, in conjunction with a method such as the elbow technique [Marutho et al., 2018].

In NLP, unsupervised methods are not as widely used as supervised. This is partially because supervised methods are usually more appropriate for NLP tasks where you know labels for the data, e.g. PoS tagging. There are, however, some key uses of unsupervised learning, such as web spam detection [Urvoy et al., 2008]. In authorship analysis, clustering has been used to group similar texts [Pillay and Solorio, 2010]. This could be useful if, for example, you didn't know how many authors there were. Mikolov et al. [2013]'s Word2Vec is a good example of an unsupervised method which has become very widely used in NLP. Topic Modelling can also be unsupervised – a model will learn the topics of a set of texts without explicitly being told what the topics are [Blei et al., 2003].

Explainability

Machine learning methods are often accused of being "black box". Decisions come out of the box but what happens inside the box remains a mystery. Sometimes we want a window into that box; we want to understand why a classifier is making its

decision [Vellido Alcacena et al., 2012].

Detecting subtle characteristics of language, such as deception, is not easy, so classifications are not likely to be particularly certain. Any system for identifying false information is therefore better thought of as advice to a user rather than as a hard ruling. The user needs to understand why the classifier has made its decision, so that they can make their own truth assessment.

Various methods can be used to explain a classifier's decision. Some classifiers provide weights for each of the features. SVMs, Logistic Regression and Random Forest classifiers all provide some form of weight for each feature. While these can give an impression of how important features are, the features themselves sometimes are not very interpretable. This is one reason for using feature engineering [Scott and Matwin, 1999]. Instead of using highly data-driven features, such as bag-of-words, we create features that are powerful at capturing elements and properties of the thing we are looking for. These features are based on past research and domain expertise.

Neural networks are particularly bad for the black box problem, trading performance for explainability. This is because they create their own features that may be more powerful than handmade features, but not as intuitive to somebody trying to interpret them. Attempts have been made to reduce this problem [Lei et al., 2016, Kshirsagar et al., 2017]. One technique is using 'Attention', which can highlight certain words that were particularly important for a model's decision [Ghaeini et al., 2018]. Some work has been critical of this approach [Serrano and Smith, 2019], and making attention more interpretable is a subject of ongoing work [Mohankumar et al., 2020].

2.1.4 Measuring Similarity of Texts

Many NLP techniques, including several of the methods we will employ later in the thesis, involve calculating the similarity between texts, represented by feature vectors. There are different ways to think about similarity, and different features and similarity metrics are better for different tasks. Sometimes it is desirable to know exactly how similar the characters in one text are to another, as is the case in plagiarism detection. In such cases, one may want to use a measure such as edit distance [Navarro, 2001]. On other occasions, it may be preferable to find out how semantically similar two texts are, e.g. for ranking search engine results. Here, embeddings and topic models may be more

useful. When performing an unsupervised analysis of text, an important consideration is the similarity metric that you use. There are many ways of calculating similarity between vectors. A survey is offered by Gomaa and Fahmy [2013]. Some of the most used similarity measures are as follows:

Jaccard Similarity Proportion of the words in a text that it has in common with another text, i.e. the intersection divided by the union.

Euclidean Distance Straight line distance between two points in vector space.

Cosine Distance Cosine of the angle between two vectors.

In natural language processing, cosine similarity is often the default as it is well suited to sparse matrices, and is not overly influenced by vector length – which makes it useful for comparing the style of authors [Evert et al., 2017], for example. However, new similarity metrics have been devised. Kusner et al. [2015] proposed Word Mover's Distance which extends a computer vision algorithm called Earth Mover's Distance [Rubner et al., 2000] to text. The word vectors making up a text are seen as point clouds and their similarity is calculated as the minimum distance needed to move the points of one text to another. This method is very computationally complex, but a simplified version which is quicker to execute was also proposed. Document Embeddings also allow comparisons to be made between texts using normal vector distance metrics such as cosine distance. This method has an advantage over comparing averaged or summed word vectors because document embeddings have some knowledge of word order in the text. De Boom et al. [2015] compared various methods for comparing the similarity of texts in vector space models. They found that none of the embedding-based methods outperformed tf-idf for long texts. Tf-idf falls apart when texts are short and there is little overlap in words between texts.

2.1.5 Stylometry

Stylometry is an area of NLP that looks at authorial style [Neal et al., 2017]. Unlike other areas of NLP, such as topic modelling or sentiment analysis, the focus is on the style of the text rather than the content. This means looking at features of the author's language that may be subconscious and separate from the topic and content of the text.

Stylometry is a task with very close links to deception detection, where detecting the underlying features of an author is important, so it is also highly relevant for work in false information.

There are several main tasks in stylometry, described by Neal et al. [2017] as:

Authorship Attribution is the task of predicting the true author of a document [Jockers and Witten, 2010]. A classic example would be attributing a historical play to a playwright of the era.

Authorship Verification is usually a binary decision as to whether or not two texts were written by the same author [Halvani et al., 2016].

Author Profiling involves predicting certain attributes of authors of a text such as age and gender [Huang et al., 2014, Reddy et al., 2016]. Author profiling has become an interesting challenge in online environments, as any user may be masquerading as somebody they are not.

Stylochronometry is the analysis of Linguistic Style over time [Stamou, 2007]. More detail on stylochronometry is provided in Section 2.3.3

Many authorship tasks can in some way be more broadly described as similarity tasks. A very important part of the field is being able to tell how similar two pieces of text are. One of the most important similarity measures that is widely used in stylometry is Burrows' Delta [Burrows, 2002]. Other similarity metrics have already been discussed in Section 2.1.4.

The features used in stylometry fall into five categories, similar to those in general NLP: Lexical, Syntactic, Semantic, Structural, and content-specific features. Character ngrams and function words are both features that are commonly used in stylometry due to their simplicity and effectiveness in many tasks. Syntactic features such as part-of-speech tags and grammatical trees also provide features that are separated from content. These features can either be profile based, where they are calculated per author, or instance based, where they are calculated per document. Grieve [2007] provided an empirical comparison of different authorship features.

Much modern stylometry research makes use of machine learning for classifying texts into different groups, for example based on the author or an author trait.

Unsupervised clustering algorithms have been used to much success [Pillay and Solorio, 2010], as have supervised classification algorithms such as SVMs [Zheng et al., 2006].

2.1.6 Ethics in NLP

Ethics is a very important aspect of NLP that has been somewhat rejected in the past. Recently, much more work has focused on this subject, looking at the ethical considerations that must be borne in mind when building NLP systems [Hovy and Spruit, 2016, Tsvetkov et al., 2018, Blodgett et al., 2020].

A major problem with many NLP systems is algorithmic bias. Biases in the data used to train systems reflect the human biases existent in society. This means that decisions made using these systems are not objective, as many may assume. For example, word embedding systems have been shown to display biases based on gender [Bolukbasi et al., 2016] and race [Manzini et al., 2019]. This is shown by the relationships between vectors. "Doctor" has the same position relative to "man" as "nurse" does to "woman". Along similarly problematic lines, "black" relates to "criminal" in the same way "caucasian" does to "police".

Biases within the dataset may be down to societal biases, for example that men get more news coverage than women [Jia et al., 2015] and dominate Twitter conversations [Garcia et al., 2014]. They may also be down to sampling bias. For example, Twitter and Reddit, two very popular data sources, only represent certain demographics: for example, people from western, industrialised countries [Henrich et al., 2010]. One cannot necessarily make general claims about people based on evidence from a limited range of sources and demographics.

One significant concern is the potential consequences of NLP systems. For example, some systems may perform worse for certain demographics of people because of biases in the dataset. Hovy and Søgaard [2015] showed that tagging systems were less effective for the language of younger people because of the data they were trained on. This is particularly problematic when it leads to situations where NLP systems work less well for minorities [Jørgensen et al., 2015]. Davidson et al. [2019], for example, showed that African-American English was more likely to be detected as abusive by various systems. If these systems were used in the wild, it would negatively affect African-American individuals. For an example of this problem relating to false information, one

may wonder what would happen if NLP fake news detection systems became widely used. Would this allow for malicious actors to build their own, censorious systems which may be legitimised by previous ones?

There is also the problem of unintended dual uses of NLP systems. For example, could authorship attribution techniques be used to identify anonymous dissenters in oppressive regimes? This issue is particularly relevant in false information. NLP systems can be used to generate, just as much as detect, fake content [Hovy, 2016]. Research into detection may make the generation even better.

There are also various ethical issues with the construction of datasets [Mieskes, 2017]. One is traceability [Couillault et al., 2014]: can individuals be identified based on information in the corpus? This is a bigger issue than it used to be as more and more work in NLP uses social media data made up of many "normal" individuals. This is an especially pertinent issue for tasks such as author profiling. There are also various issues surrounding whether or not data can be considered public. In the past anything publicly available without signing in was considered fair game without the need for getting permission [Seale et al., 2010]. Certain criticisms have arisen of this, for example that what users write in a certain community was only intended to be read by that community, and should therefore be respected as not public. Analysis of large datasets may be seen as less problematic if the analysis is purely aggregational, rather than looking at individuals [Rivers and Lewis, 2014].

The ethical problems here are not to suggest that we should avoid doing any research with datasets that may be at all biased. Rather, the point is to highlight how important it is for researchers to consider these issues when designing systems and reporting findings. This is especially crucial in a relatively new and fast moving area such as false information research, to ensure that ethics does not get left in the dust of rapid progress.

2.1.7 NLP Summary

In this section we have summarised several areas of Natural Language Processing to provide useful context for this work. We described various representations that can be used to represent texts. Next we mentioned some machine learning methods that are common in NLP work. Following that we talked about different ways of measuring

the similarities of texts. Finally we described the area of Stylometry, which is very relevant to this work. Much of the past work in Natural Language Processing has treated language as static. Increasingly this is not the case. In Section 2.3 we highlight some methods that analyse language over time.

2.2 Corpus Linguistics

Corpus Linguistics is an area of linguistics that focuses on the study of large sets of machine-readable texts, known as **corpora**. McEnery and Hardie [2011]'s book on the subject, which forms the basis of this section, outlined two generalisations about Corpus Linguistics to go towards forming a definition. The first was as follows:

> "We could reasonably define corpus linguistics as dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions. The set of texts or *corpus* dealt with is usually of a size which defies analysis by hand and eye alone within any reasonable timeframe." - McEnery and Hardie [2011], pg. 1

This provides a working definition of corpus linguistics that we will follow through this thesis. The key points are that corpus linguistics involves looking at large corpora, to answer research questions. The large size of these corpora motivates the use of machines to read the text and leads on to the second generalisation.

> "Corpora are invariably exploited using tools which allow users to search through them rapidly and reliably." - McEnery and Hardie [2011], pg. 2

This gives an idea of the methodologies involved in corpus linguistics. Methods and tools are created that allow human analysts to work with large datasets that would have previously been infeasible. These tools provide functionality such as analysing word frequencies, or creating concordances.

Corpus Linguistics as a field began in the 1960s, and accelerated as more available computation made large corpora increasingly feasible. The Survey of English Usage (SEU), founded by Randall Quirk in 1959, was the first European Research centre to conduct research using corpora. Over the following decades, various research

groups created corpora, such as the London-Lund Corpus[2] and the British National Corpus[3], among others. At the same time, computational methods were becoming more advanced, allowing the creation and processing of larger corpora. For example, Part-of-Speech taggers, such as CLAWS [Garside, 1987], removed the need for linguists to manually assign grammatical tags to millions of words of text. More recently, tools such as Sketch Engine[4] [Kilgarriff et al., 2004], WMatrix[5] Rayson [2008], and CQPweb [Hardie, 2012] have been developed that have made it easier for linguists to perform analyses.

While Corpus Linguistics is clearly related to Computational Linguistics/NLP, there is a key difference that sets them apart. McEnery and Hardie [2011] lay out this difference.

> "Corpus linguistics is ultimately about *finding out about the nature and usage of language*. While computational linguistics may also be concerned with modelling the nature of language computationally, it is *in addition* focused on solving *technical problems involving language*." - McEnery and Hardie [2011], pg. 228

Despite this difference, both fields have something to give the other. Computational linguistics introduces new methods for processing large quantities of data, providing tools for the gathering and analysis of larger, more complex corpora. Corpus linguistics, on the other hand, provides well constructed datasets to be used by computational methods, and means of analysis that provide more easily understandable insights into language than the often black-box methods used in NLP.

In this section, we will explain some key concepts from Corpus Linguistics that are relevant to the work in this thesis. This predominantly includes aspects of corpus creation, as well as common methods for analysis. Corpus Linguistic methods are well suited to the work in this thesis, as they can provide more easily interpretable results than those obtained with more computationally complex NLP methods. While some works have used Corpus Linguistics to study subjects relating to this thesis, such as

[2] `http://korpus.uib.no/icame/manuals/LONDLUND/INDEX.HTM`
[3] `http://www.natcorp.ox.ac.uk/`
[4] `http://www.sketchengine.eu/`
[5] `http://ucrel.lancs.ac.uk/wmatrix/`

false information [Dance, 2019] and deception [Gillings, 2021], the key benefit we seek to draw from CL is this added level of explainability.

2.2.1 Analysing Corpora

Corpus Linguistics centres around the building and analysis of corpora. Corpora are built to be representative of the type of language being studied, and can contain spoken or written language data from many genres and sources. Corpora can be built, for example, out of general language from a certain era[6], or more constrained language, such as that of parliamentary debates[7]. These corpora may then be annotated to mark up the text with additional information, such as PoS or Semantic tags (see Section 2.1.1).

The analysis of corpora usually involves the use of software tools that allow quick analysis of a larger amount of data. While there are many methods and techniques employed in the field, we will only describe a handful of particularly significant ones which apply to the work in this thesis.

Word Frequencies and Keywords

The most basic way to make comparisons within (or between) corpora is to report **word frequencies**: the number of times a word appears in a corpus. This allows simple comparisons, such as demonstrating that one word is more frequent than another, or that it is more frequent in a certain corpus. One flaw with this method is that it does not account for the length of the corpus. For this reason, it is common to present **normalised frequencies** alongside raw frequencies. The normalised frequency is usually calculated by dividing the raw frequency by total number of words in the corpus. This can be multiplied by any arbitrary value of n to tell us the number of times a word appears per n words. Normalised frequencies are more useful for comparing values between corpora, as they scale based on text length. It is still worth looking at raw frequencies in conjunction, however, as normalised frequencies can be deceiving, for example if the word only appears a very small number of times.

While frequencies are useful for comparing corpora, one problem with them is that they do not give us any idea if the difference is significant, or merely down to

[6] https://varieng.helsinki.fi/CoRD/corpora/LOB/
[7] https://www.english-corpora.org/hansard/

6 occurrences.			Extend context
rising middle class . We know that	America	thrives when every person can find	1 More | Full
it . But we reject the belief that	America	must choose between caring for the	2 More | Full
long and sometimes difficult . But	America	can not resist this transition , we	3 More | Full
durably lift suspicion and fear .	America	will remain the anchor of strong al	4 More | Full
, hopeful immigrants who still see	America	as a land of opportunity -- (appla	5 More | Full
rever bless these United States of	America	.	6 More | Full

Figure 2.1: Concordance produced by Wmatrix [Rayson, 2008] for the word "America" in Barack Obama's presidential inauguration speech.

coincidence. To solve this problem, we have to turn to statistical tests. Words that are found to be significantly more frequent in one corpus than another are considered to be **keywords**. This method works equally well for any type of token, so you could also calculate key PoS or semantic tags [Rayson, 2008].

Three main significance functions have been used for this purpose in corpus-linguistics: chi-squared test, t test, and log-likelihood [Dunning, 1993]. The Chi-squared test relies on the assumption of data being normally distributed, which makes it often inappropriate for text data (which is rarely normally distributed). For this reason, the log-likelihood statistic is often preferred in corpus-linguistics today, as it does not assume the normality of data. For more information on significance testing in corpus linguistics, various books have been written about the use of statistics within the field [Brezina, 2018a, Gries, 2013].

One other statistic worth mentioning is log-ratio[8]. This is an effect size rather than significance metric. It tells us how many times more likely a word is in one corpus than another. A log ratio of 1 means a word it twice as likely (2^1), and a log ratio of 2 is four times as likely (2^2). This is more intuitively readable than a significance score, and can also be used to compare words between multiple corpora of different lengths, which is not possible with log-likelihood.

Concordances

Concordancers are one of the most universal methods in corpus linguistics. These pieces of software allow all instances of a given string of text to be printed within a context window. Often these strings are words, but they could just as easily be a suffix or a multi-word expression. The process of displaying words within a context window is called "keyword in context" (KWIC). KWIC concordances are produced using software tools

[8] http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/

such a WMatrix [Rayson, 2008] and Sketch Engine [Kilgarriff et al., 2004]. Figure 2.1 shows an example of a KWIC concordance produced by Wmatrix.

Concordances are a useful way of looking at language features of interest within a corpus. This analysis is usually fairly qualitative, with the analyst looking for patterns that make sense to them, or seem to answer one of their research questions. As with simple word frequencies, concordances by themselves do not show the significance of words appearing together. To do this, we would instead look at collocations.

Collocations

Collocates tell us similar things to concordances in a more quantitative way. They rely on the idea discussed by Firth [1957], that the meanings of words do not simply depend on the word in isolation, but also on the other words with which they co-occur. This is the same idea that underpins more recent NLP work on word embeddings [Mikolov et al., 2013]. The collocates of a word are the set of words with which the word frequently co-occurs.

The size of the window in which the words co-occur can be varied. For example, one may only consider sequences of words (ngrams) [Harris, 2006], or one could consider words that appear within a window of n words, regardless of ordering [Sinclair et al., 2004]. Window size will change the types of collocates that are captured. A larger window will capture longer-distance collocates, which may link to more semantic relationships, while a smaller window will capture more grammatical relationships.

Because many words co-occur with many other words, only words that pass through a significance filter are usually considered as collocates. This means that collocates are pairs of words that co-occur significantly more frequently than those words with other words. Two common significance metrics used for the purpose of identifying collocates are mutual information and log-likelihood. The chosen significance measure can dramatically change the collocates found, as each metric will rank different collocates more highly. For example, Mutual Information tends to highlight rarer word combinations, and log-likelihood more common. Because of this, it can be useful to look at both when performing an analysis. For more information on the parameters and statistics behind collocation, various works have been produced that look at this in much more detail [Church and Hanks, 1989, Dunning, 1993, Evert, 2005].

2.3 Language Change

Often in NLP, data is viewed as somewhat static and changes over time are not considered. For many NLP tasks, such as machine translation, it is not as important that you take into account time. But language, particularly on the web, is fast moving and dynamic so it is useful to understand how changes over time manifest.

In this section, we will discuss some of the work within NLP that looks at language change. More detail about some specific methods will be provided in Chapter 4. We will touch upon some applications of language change techniques on online communities in this section, however this subject will mainly be addressed in Section 2.4 and Chapter 7.

2.3.1 Computational Approaches to Language Change

Studying language change has become a popular topic of research in NLP and Computational Linguistics. Many traditional techniques are predominantly designed to treat text as static. Creating new methods, or adapting old methods, for dealing with diachronic data is an important research challenge. Much of the work has focused on the changes in word usage over time. In a tutorial on the subject, Eisenstein [2019] provided a look at various methods for studying language change.

Linguistic Innovations

A common task which requires the analysis of language over time is the identification of linguistic innovations. In the past, this is a problem that has been looked at on a longer term scale, e.g. over decades or centuries, but with the advent of the internet, online communities became interesting case studies to observe these types of change over shorter timescales [Bucholtz, 1999].

Some works have sought to analyse the way linguistic innovations spread, and the reasons behind them. Altmann et al. [2011] looked at social dissemination in online groups, and discuss the effect of a word's niche on its success. Another interesting example from Garley and Hockenmaier [2012], studied the introduction of English loanwords in a German language music forum. Del Tredici and Fernández [2018] looked at how central community members facilitate the spread of new terms. Del Tredici et al. [2019] then looked at short term meaning shift in an online football

community, using distributional techniques similar to Kim et al. [2014]. They showed how certain common words changed their usage as they picked up novel meanings within the community. In a slight departure from words, Tsur and Rappoport [2015] looked at Twitter hashtags as neologisms.

Other works have taken a broader view, rather than focusing on niche communities. Grieve et al. [2017] looked at lexical emergence in American English by looking at Twitter. Stewart and Eisenstein [2018] observed the change in usage of non-standard words on Reddit, finding that dissemination across many linguistic context often led to a word's survival. In a study that covered Twitter and Reddit, Kershaw et al. [2017] found that community structure had little effect on the propagation of out-of-vocabulary word adoption.

Semantic Shift

In recent years, a substantial amount of work has been put into researching Semantic Shift. Kutuzov et al. [2018] provide a survey of recent work. Semantic Shift is the way that the meaning of words change over time and can be both a result of a linguistic drift or a cultural shift [Hamilton et al., 2016c]. Over the past few years, the majority of approaches to this task have used distributional word representations, e.g. Baroni et al. [2014]. Kulkarni et al. [2015] provides a comparison of distributional and frequentist approaches. Following on from this Giulianelli et al. [2020] presented the first work using contextualised BERT representations to observe lexical change over time.

One challenge with using word embeddings is that if you train two embedding models on the same data, the numerical vectors will be different for the same words. This means that you cannot simply detect change by looking at the cosine distance between two vectors. Kulkarni et al. [2015] suggested aligning the embedding models by performing linear transformations that preserve the models' structure but make them comparable. There are other methods of comparing embeddings, such as that proposed by Eger and Mehler [2016], who analysed the position of words with respect to other words, and of Kim et al. [2014], who incrementally trained embeddings for each year using the previous year's values for initialisation. Yao et al. [2018] showed that you could train word embeddings jointly across multiple time periods, which enforced the alignment of the embedding models.

Recently, Gonen et al. [2020] proposed a system to look at usage change between corpora which does not require alignment. The method achieves this by finding the nearest neighbours (using cosine similarity of vectors) of a word in both corpora. The overlap between these lists of neighbours suggests how much a word changes between these corpora. This method does not exclusively have to be used to compare different time periods: it could just as well compare the language usage between geographical regions or age demographics.

Diachronic word embeddings exhibit interesting properties similar to normal word embeddings. For example, one can witness semantic relations such as how the word 'Obama' in 2008 is very close to the word 'Bush' in 2004, an example of which was shown by Yao et al. [2018].

Comparing Differences Over Time

Often language change research involves comparing the linguistic differences of different people over time. An example of a task that demonstrates some of these methods well is that of measuring polarisation and partisanship in politics [Iyengar et al., 2012, Monroe et al., 2017]. Demszky et al. [2019] looked at lexical and topic features, and showed that polarisation has increased with regards to mass-shootings.

One method that has been used to make comparisons over time is classification accuracy. Peterson and Spirling [2018] used this metric to chart polarisation in UK politics. A similar method was employed by Underwood et al. [2018], who looked at gender in English language fiction, showing that the distinction between genders of both characters and authors have become blurred over time.

Information theoretic techniques can also be used for comparing different users or corpora over time. Danescu-Niculescu-Mizil et al. [2013] used cross-entropy to compare users to a community over time, while Barron et al. [2018] used Kullback Leibler (KL) Divergence [Kullback and Leibler, 1951] to study the language of the French Revolution.

Linguistic Influence

Some works have looked at linguistic influence within communities. Influence can take many forms. It could mean influence between geographical locations [Eisenstein et al.,

2014], or looking at the types of users that adopt new conventions in social media [Kooti et al., 2012].

Various works have looked at influence in different ways on social media. Danescu-Niculescu-Mizil et al. [2011] observed the phenomenon of linguistic style accommodation [Niederhoffer and Pennebaker, 2002] in Twitter conversations, creating a framework to measure accommodation. Goel et al. [2016] examined language change on Twitter, looking at the spread of novel spellings and abbreviations, and finding that certain social connections seem to yield more influence than others. Leskovec et al. [2009] tracked memes through the web, observing how phrases were picked up by the news cycle.

Not all research on this topic has looked at social media text, however. Some work has looked at academic papers, such as Gerow et al. [2018], who measured influence in academic papers going back more than 100 years. They showed how certain factors boost the influence of academics, by looking at which works have high influence, but a low number of citations. In doing this, they built on the work of Gerrish and Blei [2010], who created a model for predicting the impact of academic publications. Another work in the non-social-media space is that of Guo et al. [2015], who produced a model for looking at social influence and linguistic accommodation in transcripts of conversations.

Different Time Ranges

Methods looking at language change can focus on text over different time ranges. Some techniques look at the change of language over decades or even centuries, while others over years [Gerow et al., 2018]. On social media, as there is a lot of data spanning a shorter length of time, works often focus on time ranges of months [Kooti et al., 2012, Goel et al., 2016, Stewart and Eisenstein, 2018], observing how language changes during a community's lifetime. When looking at a certain event, one may end up looking at days [Leskovec et al., 2009]. Other works have gone even finer still, looking at language change over a matter of minutes or hours [Danescu-Niculescu-Mizil et al., 2011, Golder and Macy, 2011, Doyle and Frank, 2015].

Syntactic Change

Most of the works listed so far have looked at lexical change. Some research has focused instead on syntactic change, which tells us less about topics and terminology, and more about how style and grammar change over time. For example, Perek [2014] used a distributional measure of semantic similarity to study syntactic productivity over time. Degaetano-Ortlieb and Teich [2018] introduced an approach for identifying features involved in language change, and identifying periods in a corpus of academic text. They did this using KL Divergence [Kullback and Leibler, 1951] with part-of-speech trigrams. Because syntactic change is more linked to style, some other methods looking at this will be described in Section 2.3.3.

Topical Change

Another way of looking at change over time involves looking at changing topics [Wang and McCallum, 2006, Blei et al., 2003, Hall et al., 2008]. Blei and Lafferty [2006] introduce the idea of dynamic topic models, which allow the topical analysis of large diachronic corpora. Hall et al. [2008] used LDA [Blei et al., 2003] to track the development of ideas across multiple conferences in the ACL Anthology, comparing topic distributions using Jensen-Shannon divergence. Other works have looked at how topics change in conversation [Nguyen et al., 2014], and used topic to analyse polarisation [Demszky et al., 2019].

Epoch Detection

Another interesting task relating to language change is Epoch Detection, which we will define as automatic methods for identifying linguistic periods within a corpus. In the past, this has been done manually, often by arbitrarily splitting data into regular time intervals. Gries and Hilpert [2008] suggested a solution to this problem: a clustering-based method used to split a corpus into stages, based on a similarity metric and a set of hand-crafted features. van Hulle and Kestemont [2016] used stylometry to identify stages in the language of Samuel Beckett. We have already discussed the work of Degaetano-Ortlieb and Teich [2018], who used KL Divergence to periodise a corpus of academic text. Statistical tests have also been employed to identify epochs [Popescu and Strapparava, 2013, 2014].

2.3.2 Corpus Linguistic Approaches to Language Change

The field of Corpus Linguistics has long been interested in the way language changes over time. So it is no surprise that corpus linguists have spent years creating diachronic corpora (see Section 2.3.4), and using them to perform analyses. Several works have surveyed these approaches [Hilpert and Gries, 2016, Brezina, 2018b, Kytö, 2011]. Many entire books have been written on this subject [Leech et al., 2009, Whitt, 2018], so needless to say this section will only provide a brief overview of these works. Though there is some overlap with NLP/computational methods, we have kept this separate as NLP and corpus linguistics approach the same problems from different perspectives.

Hilpert and Gries [2016] suggest the following questions which these methods typically seek to answer. As quoted from the paper:

- When and how does a given change happen?
- Can a process of change be broken down into separate phases?
- Do formal and functional characteristics of a linguistic form change in lock-step or independently from one another?
- What are the factors that drive a change, what is their relative importance, and how do they change over time?
- How do cases of language variation in the past compare to variation in the present?

A large amount of the corpus research looking at language change is focused on general change in a language, usually English. This contrasts to some of the work in NLP, particularly work looking at social media, which is often interested in the language of certain groups or communities over time [Del Tredici et al., 2019, Demszky et al., 2019]. It is not true, however, that corpus research is exclusively interested in general change. There is work such as McEnery and Baker [2016] that looks at the language surrounding particular social phenomena to learn more about historical perceptions. Such work overlaps with other fields of research in the humanities.

This analysis can be done over both long [Curzan, 2009] and short [Baker, 2011, Mair, 2009] time periods. Work has looked at both historical [Kytö, 2011], and more contemporary corpora [Baker, 2011].

In terms of methods, there have been various approaches. On a basic level, many of these techniques involve corpus comparisons [Hilpert and Gries, 2008, Gablasova et al., 2017]. Different features have been analysed over time, such as register [Biber and Gray, 2011], grammar [Leech et al., 2009], and style [Leech et al., 2012]. Gries and Hilpert [2008] introduced a method, Variability-based Neighbour Clustering (VNC), which finds stages in diachronic corpora in an unsupervised fashion. McEnery et al. [2019] built the peaks and troughs method [Gabrielatos et al., 2012] to plot the fluctuation of a word's usage over time based on its collocates.

2.3.3 Stylochronometry

Stylochronometry is a sub-field of stylometry concerned with the analysis of changes in authorship style over time. In a survey of approaches to Stylometry up to 2007, Stamou [2007] identified three common tasks: identifying stylistic development of an author, finding the sequence of composition for the works of an author, and relative dating of an author's works.

An early work in stylometry was carried out by Forsyth [1999] who studied the changing language style of American poet William Yeats. The author used a method of finding distinctive substrings which determined whether a poem was written by "Young Yeats" or "Old Yeats". Pennebaker and Stone [2003] took a psychological approach and examined the texts of various authors over time to see if the language of individuals changes over their lifespan. They identified several differences, such as how older users used fewer self-references and a general increase in cognitive complexity. Can and Patton [2004] looked at the changing style of two Turkish authors over time. They looked at the change of three linguistic markers: frequency of word lengths for text and vocabulary, and usage of the most frequent words. They found that words were longer in newer texts. Hoover [2007] used various methods including word frequencies to analyse the style change of author Henry James. More recently, Klaussner and Vogel [2015] looked at two authors (Henry James and Mark Twain), comparing their language over their careers. They used a regression to predict the year of a given text, with some success. The authors expanded on their regression work in a subsequent paper [Klaussner and Vogel, 2018b].

In a slightly different setting from literary text, Degaetano-Ortlieb [2018] examined

historical court proceedings from the Old Bailey. Relative entropy was used to find stylistic variation between different groups, focusing on gender and social class.

Fifield et al. [2015] used clustering to perform unsupervised authorship attribution over time. Data was split into sequential chunks and was clustered. This was used to try and resolve the author of an ancient poem.

One key issue when performing temporal analysis of text is finding good quality datasets that cover a given time span, while containing enough text at each time step. Much work has been concerned on dating historical texts, such as those of Shakespeare, Marlowe, and Plato. One limitation has been that research has, for the most part, only looked at one or two authors. Klaussner and Vogel [2018a] aimed to address this issue by creating a corpus designed for analysing the language of multiple late 19th and early 20th century authors.

2.3.4 Diachronic Corpora

A diachronic corpus is a corpus that contains texts, from a certain time range, which represent a sample of linguistic development. Within Corpus Linguistics, various diachronic corpora have been created. Some of these aim to capture general language change, e.g. the Corpus of Historical American English (COHA) [Davies, 2012], while others focus on a specific source or genre of text, e.g. UK parliamentary debates[9].

In Computer Science and NLP, there is sometimes more of a focus on quantity over quality. Unlike in corpus linguistics, where corpora are carefully constructed to ensure representativeness, datasets in NLP often involve collecting all available data from easily accessible sources. Some of the problems this can cause have already been discussed in Section 2.1.6. Large datasets commonly used in NLP are often gathered from social networks such as Twitter and Reddit, which have APIs that allow the scraping of pretty much everything ever published on them. There are also non-social media based datasets which can be used to perform diachronic analysis. Google Books Ngrams has been used for analyses of language over time [Michel et al., 2011], and is used for training language models for looking at semantic change [Kulkarni et al., 2015], amongst other things. This dataset is limited because it does not contain full texts, but rather ngrams, as the name suggests. Another dataset used in NLP for language change

[9] `https://www.english-corpora.org/hansard/`

is the Gigaword corpus[10] [Kutuzov et al., 2017], which is composed of newswire text. Some NLP works have also used the corpora created by linguists, such as Eger and Mehler [2016] and Hamilton et al. [2016b], who made use of COHA.

In NLP, as in Corpus Linguistics, there are also various more tailored corpora, which focus on more limited time frames, genres, or communities. For example, some work looks at change during a specific period of time, such as the French Revolution [Barron et al., 2018]. Genres such as academic text may be explored with corpora of academic papers, for example the ACL Anthology [Hall et al., 2008] or Royal Society corpus [Fischer et al., 2020]. Online text can also be scraped from a specific community [Del Tredici et al., 2019], or genre, e.g. reviews[11] [Kulkarni et al., 2015].

2.3.5 Language Change Summary

Over the course of this section, we have introduced the problem of language change, and described different approaches to its study. To date, much of this work has focused on observing linguistic innovations, and words with changing meaning. These works give more of an impression of general language change, as opposed to the changes in individuals and groups. Some work has been done on language influence within groups, and comparing individuals to the community, however communities often contain sub-groups of users within them, for example, trolls and moderators. More work could be done into the comparison of such sub-groups. In false information related communities, there may exist sub-groups relating to belief. Observing how these groups change over time, and influence each other, will lend key insight into the way false information is discussed and spread. There also needs to be more work looking at language change in the specific context of false information.

2.4 Online Communities

The prevalence of social media, and the ability to scrape huge amounts of natural language data from platforms such as Twitter and Reddit, has led to it becoming an area of focus in Natural Language Processing. Traditionally much research on

[10] https://catalog.ldc.upenn.edu/LDC2003T05
[11] https://snap.stanford.edu/data/web-Movies.html

the language of social media has treated the data as static, not taking changes over time into account. However, if we wish to understand how information (including false information) spreads in communities, we must consider the effect of time on the language of users. This includes looking at linguistic influence exerted between users, and examining the changing norms of language within and between communities. This section will describe recent Social Media and Natural Language Processing research on online communities. A lot has been done in this area, so we will focus on work which relates to the language and behaviour of users in communities, as well as work relating to false information and language change. Some of the work discussed here will overlap with Section 2.3.1, though here we will not be exclusively looking at NLP methods. There will also be overlap with Section 2.5.2, which discusses social network approaches to fake news detection.

Many of the works on online communities are focused on the networks linking users. Community detection algorithms [e.g. Rosvall and Bergstrom, 2008, Blondel et al., 2008, Ronhovde and Nussinov, 2009] use the connections between nodes in a social network to detect communities. Lancichinetti and Fortunato [2009] compared various approaches, though they did not account for hierarchical community structures or networks with no communities. Azaouzi et al. [2019] provides a recent survey of the state-of-the-art for community detection. On a theoretical level, most of these techniques boil down to a graph partitioning problem [Fortunato, 2010, Rhouma and Romdhane, 2014]. While it is useful to be aware of this research area, our work is more interested in linguistic relationships between users.

Of the major social media platforms, Reddit is of most interest for our work, as subreddits more closely resemble communities of like-minded individuals than more general social media platforms such as Twitter and Facebook. Medvedev et al. [2019] provides a survey of current research making use of Reddit data. They split work on Reddit into two broad categories: post-level, and user-level techniques. We will discuss some of this work, focusing primarily on work that considers the language features of users.

Post-level techniques involve using posts as the unit of analysis, often making predictions about them. A common task is predicting the popularity of posts [Horne et al., 2017, Fang et al., 2016, Jaech et al., 2015]. This can be done based on structural

features, language usage, or author attributes. The task can be set up as a prediction task, with the user generated scores of posts used as labels to predict. Horne et al. [2017] performed an analysis of post popularity, and found certain global features, such as emotional content, were predictors of popularity. Other features varied greatly across communities.

User-level techniques look at the users who post in the community, and the networks that connect them. This includes looking at posting activity. Glenski et al. [2017] looked at browsing and voting behaviour of users on Reddit. This helped paint a picture of how users use the website. For example, they showed that most users vote without reading an article. Users tended to not interact with posts, mainly just reading headlines. They found that they could predict the interactions of users with relatively simple models [Glenski and Weninger, 2017].

Tran and Ostendorf [2016] compared style and content based features for predicting reception to posts in a community. They found style features were a better indicator of the community than content features, though their style features were slightly content dependent, being a combination of the most frequent words and PoS tags. Users whose posts were similar to the style of the community were usually more popular.

Trolling and hate-speech are other aspects of online communities that have been well studied. Some works have analysed abusive behaviour in online communities [Chatzakou et al., 2017]. Chandrasekharan et al. [2017] looked at the aftermath of two subreddits being banned in 2015. Their findings suggested that the ban resulted in numerous users leaving the platform, while those that stayed reduced their hate speech usage. Mojica de la Vega and Ng [2018] built and annotated a dataset containing examples of trolling, labelled from both the troll and recipient's perspectives.

Plenty of work has been done outside of Reddit, too. A huge amount of research has focused on Twitter, though as we mentioned earlier, it feels like less of a coherent community than smaller groupings such as subreddits. Platforms such as Facebook [Lambiotte and Kosinski, 2014] and MySpace [Lee et al., 2012] have also been studied, and allow insight into a slightly more personal type of online discussion. YouTube is another platform on which community dynamics have been looked at [Szabo and Huberman, 2010].

We have already described several works in Section 2.3.1 which looked at the

emergence of new terms and conventions in social media. Some of these looked at linguistic developments within niche communities, such as Del Tredici et al. [2019], who looked at how words gained new meanings within a football team's fan community. Another work that looked at the emergence of terms in a community was Garley and Hockenmaier [2012], who track the usage of English loan-words in an online music community.

Membership of Communities

Hamilton et al. [2017] looked at loyalty in multiple Reddit-based communities. They found that loyal users (users that stay on the forum over a long period of time) have certain shared features across communities. Loyal users use similar language signals, and engage with less popular content, suggesting they may act as trend-setters. They found that a user's loyalty could be predicted based on early interactions.

Along similar lines, Newell et al. [2016] observed the migration of users following the closure of several communities on Reddit in 2015. Their findings suggested that no mass exodus away from Reddit occurred. Though some users did migrate to other sites, they found that no other platform could compete when it came to niche communities. Zhang et al. [2017] looked at the distinctiveness and dynamism of Reddit communities. Particularly, they found that communities with distinctive and dynamic identities more successfully retained users, but that this distinct identity could also make it harder for newcomers to join.

Other works looked at the growth of Reddit communities. Kairam et al. [2012] investigated what makes groups grow and last, finding that communities which recruit members from their existing network grow quicker but do not grow as large in the long run. Tan [2018] proposed a method of examining how communities are formed using genealogy graphs. They demonstrated that one can predict a community's growth using its origin.

An interesting feature of Reddit is the existence of highly related communities that exist parallel to one another. Certain subjects may have multiple communities dedicated to them, and certain subreddits may split off from existing ones. Hessel et al. [2016] looked at such examples, by identifying communities that have affixed versions of another subreddit's name (e.g. "space" and "spaceporn"). They found that users who

split off from an existing community to a new spin-off community were still more active in the original community, suggesting they do not entirely migrate.

Several works have looked at the way the communities interact with each other, as well as with the outside world. Kumar et al. [2018] looked at conflicts between subreddits, finding that a small number of communities are responsible for initiating conflicts. Active community members begin these conflicts, but less active users often participate in the conflict itself. Conflicts between users had lasting effects, such as lingering bad behaviour, in the targeted communities. They also built a model to try and predict these conflicts before they happen.

Other works looked at how online communities can influence the web on a wider level. Moyer et al. [2015] showed that topics discussed in a popular Reddit community led to large increases in traffic to Wikipedia pages relating to those topics. Zannettou et al. [2017] found that small web communities could have a disproportionate effect on large, mainstream social media platforms.

Niche Platforms

Some works have looked beyond the mainstream, to more niche platforms. Zannettou et al. [2017] looked at the way that alternative and mainstream web communities share alternative and mainstream news sources. They primarily focused on how these communities influenced each other, using Hawkes Processes to measure the influence. While they found, perhaps unsurprisingly, that Twitter was the most influential platform, their results suggested that fringe communities on 4chan and Reddit were having a disproportionate influence on what was shared on Twitter.

In a following work Zannettou et al. [2018b] measured the propagation of memes across the web. Memes are used by fringe web communities as a means of communication, usually taking the form of images, videos, and slogans. Zannettou et al. [2018b] found that the 4chan board '/pol/' exerted substantial influence on the overall meme ecosystem and reddit community 'The_Donald' often pushed memes into more mainstream communities.

Other communities that have been researched in a similar manner include 4chan [Hine et al., 2017], Gab [Zannettou et al., 2018a], and Voat [Papasavva et al., 2021] Looking at niche communities such as these serves to increase the variety of research in the area.

Different communities exhibit different behaviours, so studying a greater range of them allows us to learn about certain groups, and also to better identify universal behaviours.

2.4.1 Language Change in Online Communities

Much work has looked at language change in online communities. As mentioned in Section 2.3.1, some works have aimed to look at the way that the language of users can influence others. For example, Danescu-Niculescu-Mizil et al. [2011] showed that, even under the constraints of Twitter, users change their language style within conversations to accommodate others. While not in the online domain, Barron et al. [2018] used LDA and Kullback-Leibler Divergence (Relative Entropy) [Kullback, 1997] to find influential speakers that shaped French Revolutionary debates. Looking at the way users influence each other can increase our understanding of the way online communities communicate; for example, which users establish norms and conventions [Kooti et al., 2012].

Other works have studied the way language change over time aligns to the language change of the community. Nguyen and P. Rosé [2011] looked at how the language of long-term users in an online community changed over time. They showed that the language of users increasingly conformed to community norms over their first year of membership before it stabilised. They performed a regression to estimate how long users had been members and their results suggested that long term users are more social and use more jargon. Danescu-Niculescu-Mizil et al. [2013] looked at the lifecycle of users on two beer reviewing Communities. By looking at the cross entropy of different user's language models compared to those of the general community, they observed that users begin dissimilar from the community, assimilate, and then drift away. They suggest that users adapt to new trends early in their lives but then become 'conservative' and will not adopt anything new, causing the community to drift away from them. This corresponds to sociological work done in offline communities, looking at how individuals' language changed over their lives [Labov, 1966, 2011]. They also show that, to some extent, one could predict if a user would leave based on early posts.

Tan and Lee [2015] built on this by looking at the difference between users who leave and those who do not in a multi-community setting. They show that they can predict somebody who will leave based on their first 50 posts. For this they used information about posting behaviour as well as linguistic data and community feedback.

Their results differed from those found in a single community which is not necessarily surprising as in a multi-community setting leaving one community does not necessarily mean leaving the overall community.

Another approach to look at language change or, more specifically, influence was by Kershaw et al. [2017] who looked at the adoption of out-of-vocabulary words on Twitter and Reddit. They looked at both macro and micro relationships: macro being between users who directly spoke to each other, and macro being between users who moved from similar geographical areas and reddit communities. They built a model which learnt thresholds at which users would be influenced (i.e. use a OOV word/phrase). Their results suggested that community structure is not hugely influential on how people are influenced.

2.4.2 Datasets

Research into online communities has made heavy use of readily available data sources, such as Twitter[12] and Reddit[13]. Twitter has a freely available API which provides access to Tweets. Until recently, to access the full "fire hose", one needed to pay a restrictively large sum of money, meaning that this option was not available to many researchers. As of January 2021, researchers are now permitted access to the full stream of data, though work prior to this date often made use of the sample. The sampling strategy of the API's sample is a black-box, and has been criticised [Morstatter et al., 2013]. However, it is generally considered "good enough", and is widely used in many studies. Twitter's terms and conditions do not allow the distribution of Twitter data, so researchers must share IDs instead of raw Tweets. This causes some problems for reproducibility, but also helps to protect the privacy of Twitter users.

Reddit is also a very popular data source [Medvedev et al., 2019]. Unlike Twitter, posts can be copied and shared (without modification), so datasets can be more easily created and made public. Reddit has a public API, but it is very restrictive for large-scale collection of text. Thankfully, Baumgartner et al. [2020] created a comprehensive dump of Reddit data, which has its own API[14], and allows the collection of mostly complete corpora for communities within Reddit. This dataset does have its limitations. Firstly

[12] `https://twitter.com`
[13] `https://reddit.com`
[14] `https://pushshift.io/api-parameters/`

it is not complete. For example, certain posts will have been removed or deleted by users or moderators. Also, as it is not gathered in a live fashion, the data is not as it was at the time of posting. Gaffney and Matias [2018] highlighted some flaws with the dataset, including wide-scale missing data. Even still, the dataset allows the collection of much more data than would be possible with the Reddit API, so it is still invaluable, with some caveats.

Other works have looked at less general online communities, such as subject specific fora [Nguyen and P. Rosé, 2011, Danescu-Niculescu-Mizil et al., 2013], or smaller platforms, such as Gab [Zannettou et al., 2018a], 4chan [Hine et al., 2017], and Voat [Papasavva et al., 2021]. Video sites such as YouTube provide interesting sources for analysis [Kleinberg et al., 2018], introducing multi-modality to works that previously focused on text [Biel and Gatica-Perez, 2010, Aran et al., 2014]. Popular social networking sites such as Facebook [Mihált z et al., 2015, Overgoor et al., 2020], WhatsApp [Sprugnoli et al., 2018, Reis et al., 2020], and Instagram [Kruk et al., 2019, Branz et al., 2020] have been studied, but are challenging due to restrictive APIs, and more serious privacy issues. For example, while Twitter profiles are generally public, Facebook profiles usually have an expectation of privacy from the perspective of users.

2.4.3 Online Communities Summary

As we have seen, there has been significant interest in the study of online communities. However, there are several areas in which this research could be advanced, which we will address in Chapters 6 and 7.

Most of the works looking at online communities have considered a given forum or subreddit as a single community, and do not attempt to look at subgroups within these communities. Studying these sub-groups would increase our understanding of the make-up of communities. In the specific case of false information, the types of user present may tell us more about how false information is spread. For example, we may wish to compare the language of believers and non-believers within a community.

Much of the work in this area looks at mainstream platforms such as Reddit and Twitter. But users from other communities, such as subject-specific fora, may use language differently compared to users on Reddit. Looking at a wider range of sources will provide useful insight into the way all kinds of community operate online.

2.5 False Information

False information seems like a simple concept: information that is verifiably false. However, in reality, False Information is an umbrella term that includes both *misinformation* and *disinformation*. In this thesis, misinformation is defined as false information that spread unintentionally, for example a rumour passed along without scrutiny. Disinformation, on the other hand is intended to deceive. These definitions are consistent with much of the literature [Hernon, 1995, Stahl, 2006, Fallis, 2015, Kumar and Shah, 2018, Guo et al., 2020], though it is worth being aware that sometimes misinformation is used as the overarching term. We are avoiding this meaning as it creates ambiguity, so "false information" will always be used in this context. *Fake news* is another popular term, often loosely defined. We will consider fake news to refer to false information that takes the format of traditional news.

Today, these concepts are more relevant than ever. False information has led to mass distrust in public health policy[15], incited various acts of violence[16], and been weaponised by world leaders as a means of discrediting the media[17]. For these reasons, it is an area which requires much research so we can better understand the way it works, with the view to help reduce its negative impact on society.

In recent years, several surveys have been carried out looking at False information. Conroy et al. [2015] provide a good overview of the different kinds of approach, as do Shu et al. [2017]. More recently Guo et al. [2020] provided a broad overview of the state of False Information detection on social media, highlighting potential future directions for the field. Oshikawa et al. [2020] performed a survey looking more specifically at Natural Language Processing approaches to Fake News detection. These surveys were all useful in forming a clear impression of all the work going on in the area of False Information.

In this section, we will consider three main categories of False Information research:

Linguistic Approaches look at the linguistic content of false information, e.g. looking at news articles, or the bodies of tweets.

Network-based Approaches use the social context of false information to make

[15] https://www.bbc.co.uk/news/blogs-trending-56526265
[16] https://www.bbc.co.uk/news/world-us-canada-40372407
[17] https://www.bbc.co.uk/news/av/world-us-canada-46175024

predictions. For example, they may look at the users spreading certain stories or claims, or the propagation patterns on social networks.

Fact Checking involves the identification and verification of claims in a document, often by comparing them to a knowledge base.

The first two of these categories are as defined by Conroy et al. [2015], while the third is kept separate because identifying and verifying facts or claims in a document is a slightly different task, though still one that is crucial in the fight against false information. These categories are only intended as a rough conceptual guide. In reality, many approaches to the problem blend elements of the three.

This section will provide necessary grounding in False Information research, which will act as context for the remainder of the Thesis. We will also review some related areas, and outline the available datasets in the field. Further related work, more specific to each chapter, will follow throughout the thesis.

2.5.1 Linguistic Approaches to Fake News Detection

By looking at linguistic features of fake news, we can see if there are certain clues hidden within text about whether something is true. This approach has the advantage over fact-checking and network approaches that it can always be applied straight away. Fact-checking and looking at sharing patterns may provide a more accurate estimation but an initial guess can be made using linguistic methods. This is important because once beliefs have been established, they can be very difficult to displace [Nickerson, 1998]. So it is important to debunk fake news as quickly as possible. Oshikawa et al. [2020] provides a survey of NLP approaches to this task.

Many works have used classification techniques to try and classify fake news articles. Some of these works make binary predictions as to whether a text is genuine or fake, as was the case for Pérez-Rosas et al. [2018], who used psycholinguistic features to try and classify fake articles. Alternatively, the problem can be treated as a multi-class prediction task [Rashkin et al., 2017, Karimi et al., 2018]. This makes the challenge of classification more difficult, but better reflects the blurry edges between genuine and fake news. Instead of classifying between genuine and fake, one might instead predict the label given to a text by a human fact checker, e.g. “mostly true”, “half

true"or "false"[18]. Nakashole and Mitchell [2014] treated the problem as a regression task, predicting a truthfulness score instead of a binary classification.

Many different feature sets have been investigated for the task of detecting false information. Rashkin et al. [2017] found several features relating to uncertainty and vagueness to be present in fake news, though they tried using these features in a classification task with limited success. If these features are useful, it would suggest a link to deception detection where these kinds of features are often used. Horne and Adali [2017] found titles to be particularly important and also suggested that fake news was more like satire than real news. Rubin et al. [2015] used Rhetorical Structure Theory and Vector Space Modelling to look at structures and similarities within fake news texts. Volkova and Jang [2018] employed psycholinguistic features to learn more about deceptive strategies involved with writing False information. Other features that have been looked at include language style [Potthast et al., 2018], sentiment [Kwon et al., 2013, Hu et al., 2014], and LDA topics [Ito et al., 2015].

Recent works have increasingly used Neural Network methods for detecting False information. The most common machine learning techniques are LSTM and CNN. Karimi et al. [2018] combined CNN and LSTM models to detect fake news. Das Bhattacharjee et al. [2017] used a semi-supervised model to predict the veracity of false information. One limitation of these deep learning approaches is that their predictions can be slightly opaque. Some works have addressed this by attempting to provide user-friendly explanations to accompany their classifications [Popat et al., 2018a].

Stance detection techniques, as used in clickbait detection, have also been used. The idea behind this is that in fake news, similar to clickbait, the body of a text will not support the headline as strongly as in genuine news. The Fake News Challenge [Pomerleau and Rao, 2017] was a shared task that invited entrants to use stance detection to identify fake news. Approaches included using Multi-Layer Perceptrons with bag-of-words [Davis and Proctor, 2017], as well as LSTMs in conjunction with statistical features [Mrowca et al., 2017].

On the web, information is not only spread through language. Other media, such as pictures or video often accompany social media posts or news articles, not to mention

[18]Examples taken from the Politifact Truth-o-meter: `https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/`

the social contexts in which they are posted or shared. This information can also be useful for verifying news [Jin et al., 2017]. The recent rise in deepfakes, especially, highlights how the problem of false information extends to video and images as well as text. Works such as those by Wang et al. [2018], Khattar et al. [2019] made use of both textual and visual features in order to detect false information. Many works have also combined text with social context features [Della Vedova et al., 2018, Castillo et al., 2011, Kwon et al., 2013].

There are several limitations of linguistic detection of False Information. Firstly, the models used to predict false information can become too topic dependant, especially if they are only trained on data from a single source. This can be addressed in future by looking at methods such as transfer learning [Ruder et al., 2019], which could help the models become more generalisable. Secondly, just looking at text means ignoring a lot of important information, such as context or visual components. This is the primary motivation for using multi-modal approaches. Thirdly, the current body of work largely ignores the differentiation between disinformation and misinformation. In most datasets, it is impossible to know whether authors believe what they are saying or intend to deceive. Understanding the effects of these aspects on the language of False Information will be very important for future work if it wants to continue using datasets where belief and intent are unknown.

Few works directly compare fake news with other types of deceptive texts. This has been done for Satire [Rashkin et al., 2017] and Clickbait [Chen et al., 2015]. By studying other sources of disinformation, we may be able to increase our understanding of how humans write deceptively, but it will also be interesting to see how these texts compare to texts in which the author is not trying to deceive but the information is still false. A more diverse range of datasets could be useful in the area of fake news detection and being aware of the linguistic differences between misinformation and disinformation could contribute to the detection of both.

2.5.2 Social Network Approaches to Fake News Detection

One way of detecting fake news is by looking at the ways in which it is spread on social media. This can provide ways of detecting fake news which do not require examining the content of an article. Network approaches to False Information detection use the

social context of posts, and also look at the network through which the information is propagated. Guo et al. [2020] split these approaches into two main subsets:

Post-based approaches look at the users sharing the information, as well as information gathered from posts, e.g. number of likes.

Propagation-based approaches use the networks through which information is spread to make predictions.

Often, these features are mixed or combined in False Information detection systems.

Many post-based approaches look at the users who share, like, and spread false information. By looking at these users, we can get an idea of how false information is spread. Tacchini et al. [2017] found information about the users liking Facebook posts useful in identifying conspiracy posts. Zubiaga and Ji [2014] found that users found the author of a post an important factor in their decision as to whether to trust it. Shu et al. [2018] looked at the different types of users that spread Fake News, trying to identify naive users, who are more likely to believe fake stories. User information can supplement linguistic features, as demonstrated by Long et al. [2017], who added profile information on top of content-based features and improved performance over state-of-the-art methods. User credibility has also been used to detect false information [Li et al., 2019a, Yang et al., 2019].

Lumezanu et al. [2012] looked at the tweeting behaviour of Twitter propagandists. Users that spread propaganda sent more tweets in shorter periods of time than other users. They also retweet frequently without posting much original content, as well as colluding with other, seemingly unrelated, accounts. It is not always necessarily humans spreading fake news either. Bots seem to be often used in the early spreading of fake news, amplifying stories and targetting high-profile users who will spread it further [Shao et al., 2018].

Similar techniques to those used in the spreading of fake news have been used by spam creators for many years. Metaxas [2010] suggest the use of anti-propaganda techniques in the recognition of web spam. Given fake news is also often labelled as propaganda, similar techniques may be applicable to fake news.

Propagation techniques look at many posts at once, considering the network of a given claim, rumour, or piece of false information. The social context of False

Information can evolve over time, as shown by Ma et al. [2015]. Several works have taken into account the structure of the propagation networks through which false information is spread [Qazvinian et al., 2011, Wu et al., 2015, Liu and Xu, 2016, Liu and Wu, 2018]. Wu et al. [2015] revealed some insight into the dynamics between "opinion leaders" and regular users in the sharing of disinformation, with rumours tending to be started by regular users and then shared by opinion leaders, with the reverse being true for genuine information.

Ratkiewicz et al. [2011] created a system for tracking political memes to detect astroturfing, smear campaigns, and other forms of misinformation. The methods of spreading fake news can also vary based on the social media platform [Mustafaraj and Metaxas, 2017]. For example, on Facebook, unlike Twitter, much of the spreading is done on private members' groups, thus making research on the matter very tricky.

More theoretical modelling techniques have also been used to try and model and prevent the spread of misinformation in social networks. This work often focuses on working out which nodes in the network should be targetted to prevent the spread of misinformation. Making agents in the network aware of the possibility of fake news [Aymanns et al., 2017] and targetting certain influential users with the truth [Budak et al., 2011, Nguyen et al., 2012] were suggested as being useful in stopping propagation. Being abstract models, these methods are not perfect but they suggest strategies that may be worth investigating in real world scenarios.

Many approaches have been a hybrid of these techniques with others, such as looking at the linguistic features of posts or users [Long et al., 2017, Kwon et al., 2013]. As mentioned in Section 2.5.1, looking at different types of information can bolster one's ability to detect false information. There is no silver-bullet feature, so a range of techniques must be employed.

2.5.3 Fact Checking

As previously mentioned, we will treat fact checking as a separate task from linguistic and network approaches. In this section, we will provide a brief overview of work in the fact checking area. For more depth, refer to Thorne and Vlachos [2018], who performed a thorough survey of work on this topic. Over the past few years, there have also been several workshops revolving around fact-checking [Thorne et al., 2018b,

2019a, Christodoulopoulos et al., 2020].

Fact checking involves analysing the coherence, logic, and context of a claim [Mantzarlis, 2015]. This is a common task in journalism, and has been done manually probably as long as the profession has existed. Automatic fact checking is desirable because it could help speed up what is quite a slow process, performed by humans. This is particularly important in a world where there is such volume of information that journalists cannot keep up.

Fact-checking often involves handling claims, and making a judgement on whether they agree with corresponding facts. The types of claim that might be checked are various, but commonly they might include numerical claims, quote verification, claims about entity or event properties, or claims regarding an entity's position on an issue. These four types of claim were featured in the HeroX fact checking challenge[19].

Many approaches use subject-predicate-object triples as input [Nakashole and Mitchell, 2014], though others have used textual claims or entire documents [Hassan et al., 2015, Vlachos and Riedel, 2015]. Using documents requires the additional step of identifying claims. Similarly to some of the linguistic methods described in Section 2.5.1, outputs can take the form of binary, or more fine-grained labels.

Fact checking techniques also require evidence. This can take the form of knowledge graphs. These can be used to either identify the graph-element that corresponds to the claim [Vlachos and Riedel, 2015, Thorne and Vlachos, 2017], or assess the probability of a claim by analysing the graph's structure [Ciampaglia et al., 2015]. The downside of the first method is that the claim must be in the knowledge graph. For the second, improbable claims will be considered false by default, but verifying the truth of improbable facts is often more important than verifying obvious facts. Some methods require evidence to be pulled from large sets of documents such as Wikipedia [Thorne et al., 2018a].

As well as using text features, fact checking can be treated as a form of Recognising Textual Entailment (RTE) [Ferreira and Vlachos, 2016]. Models predict whether a premise is for, against or observing a given claim. These methods often require the evidence corresponding to the claim to be provided. This means that sometimes a preceding step is required of retrieving the evidence from a document [Thorne et al.,

[19] `https://www.herox.com/factcheck/`

2018a].

Fact checking can also be done by finding previously fact-checked claims that match a given claim [Vlachos and Riedel, 2014]. This turns the task into a text similarity problem [Hassan et al., 2017]. Other fact-checking works have used different methods, such as using language models instead of knowledge graphs [Lee et al., 2020], or looking at the provenance of claims [Zhang et al., 2020].

There are several limitations of fact checking. Firstly, there is the inherent limitation that facts are needed to check claims against. There will always be certain pieces of evidence missing from a given knowledge graph. This may be especially true for breaking events or first reporting. Wawer et al. [2019] compared fact-checking to predicting veracity based on psycholinguistic cues, and found that most of the utterances tested simply did not provide enough information to make a fact-checking assessment.

Another possible drawback stems from fact checking's reliance on facts. If the facts were changed, the results would change. This means that there is a possibility of adversarial attacks on fact checking systems by polluting the evidence. Thorne et al. [2019b] ran an interesting shared task which involved participants creating adversarial attacks against other participants' fact-checking systems.

Another limitation is that fact checking cannot distinguish between misinformation and disinformation. This is because it simply checks if a claim is true, and does not look into the psycholinguistic features, etc, that may reveal intent.

Finally, fact checking is a very complex issue. Automatic methods currently cannot compete with human journalists and fact checkers. Notably, they cannot produce novel explanations with evidence in the same way a human can. At best they can regurgitate existing information.

2.5.4 Rumours

Rumours are awash on social media. Opinions can be swayed and false information lodged in peoples' minds by a well placed rumour. During health emergencies and disasters it is important that the general public can know what is rumour and what is fact. For this reason it is important that we are able to identify rumours. They usually fall into three main categories: True, False, and Unverified. Finding out whether a rumour is True, False, or Unverified is known as rumour verification.

The PHEME project aimed to tackle this problem by classifying rumours as well as determining their veracity [Derczynski and Bontcheva, 2014]. A method for collecting and annotating tweets relating to rumours was described by Zubiaga et al. [2015].

There are various ways of detecting rumours. One way is to look at the lifecycle of a rumour from its creation to when people stop talking about it. Zubiaga et al. [2016] showed that rumours that eventually get found to be true are resolved quicker than those that turn out to be false. Users also had a tendency to support unverified rumours. Kwon et al. [2013] looked at the temporal features of rumours, finding that rumours tended to fluctuate much more over time than other tweets. Rumours were also often shared from users with small numbers of followers to those with many. Another way to identify rumours is by looking at the responses to a rumour [Zhao et al., 2015]. This method relies on the idea that rumours will prompt more skeptical responses from other users. One can also look at rumours as sequential data like Zubiaga et al. [2017], who created a sequential classifier, looking at the features of tweets over a sequence. This approach makes use of context.

Being able to identify a rumour is one thing, but knowing if the rumour is true or credible is another [Castillo et al., 2011]. A system for determining the veracity of rumours may be useful for journalists reporting on stories, or normal people trying to get an idea of what is real. Predicting the veracity of a rumour can be done using only the content of the tweet itself [Gupta et al., 2014, Liu et al., 2015], or by looking at the responses and wider context [Derczynski et al., 2017, Kochkina et al., 2017]. False stories might have more people in the comments disagreeing with or questioning their veracity. Live systems can be created which take tweets about stories as they happen and improve system performance over time [Liu et al., 2015].

Rumours and fake news are very closely linked. Fake stories are often shared around social media to supplement rumours. While it is primarily the linguistic features of false information that are interesting for this work, it is important to understand how false stories and views of events are formed on social media as this may give a clue about how the articles are written. When rumours are spread they can be shared or reported on as fact. The individuals and news publishers that do so may believe the rumour or they may be deliberately trying to push it as true when it has been proven to be false, for

example in the notorious case of 'Pizzagate'[20], which highlights the need for us to look at belief and intent.

2.5.5 Human Approaches to Disinformation

So far, this section has focused on automated methods for looking at False Information. Naturally, these are the most prominent methods being worked on in computer science. But it is also important to be aware of some of the other methods being employed to fight false information. These methods relate more to public policy, education, and journalism. This section will provide a brief overview of some of this work.

Governments have taken an interest in false information, particularly after it was widely discussed in the context of recent political events[21][22]. The UK government discussed disinformation in their Online Harms white paper [DCMS, 2019b]. The EU also created a report on disinformation [EFAS, 2018]. These reports suggested that companies need to proactively curb disinformation by being transparent about political advertising, making it clear to users what disinformation is, and working with fact checkers, especially around election times. In their report on Disinformation [DCMS, 2019a], the UK's Department of Culture, Media, and Sport (DCMS) called for education to cultivate higher levels of digital literacy. Similar conclusions were drawn by the Cairncross Review of journalism [Cairncross, 2019], and the Children Commissioner's "Growing Up Digital" report [Children's Commissioner, 2017].

In recent years, independent fact-checkers have risen in prominence. Websites such as `snopes.com`, `politifact.com`, and `fullfact.org` all provide fact checking for news articles and claims. Some of these sites provide ratings for information, that tell a reader how true a claim is[23,24]. Mainstream media outlets such as the BBC[25], Channel 4[26], and Buzzfeed[27] have also begun to provide fact-checking. Along similar lines, there is a browser extension, NewsGuard[28], that warns users of

[20] `https://www.bbc.co.uk/news/blogs-trending-38156985`
[21] `https://www.bbc.co.uk/news/world-us-canada-37896753`
[22] `https://www.bbc.co.uk/news/blogs-trending-48356351`
[23] `https://snopes.com/fact-check-ratings/`
[24] `https://politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/`
[25] `https://bbc.co.uk/news/reality_check`
[26] `https://channel4.com/news/factcheck`
[27] `https://buzzfeed.com/uk/badge/fact-checker`
[28] `https://newsguardtech.com`

potential false information while they browse the web.

Social media companies have also taken some steps. Twitter now labels, or in extreme cases removes, potentially misleading information on their platform[29]. The company has also announced a crowd-sourced approach, in which Twitter users will be able to provide context for, or dispute, claims made on the platform[30]. The idea of this is to provide a quick response to new disinformation, but it seems potentially problematic to put this power in the hands of regular users. Facebook has been less proactive than Twitter, but also works with fact-checkers and labels misleading posts[31,32]. The social media platform-holders are in the best position to tackle the spread of misinformation, but many argue they have not done enough, or have acted too slowly[33].

Some automated methods use crowd information to aid in detection. Zhao et al. [2015] found comments to a rumour that questioned it, and used these to help identify rumours. Other systems work to automatically aid humans, such as that created by Vo and Lee [2018], which suggests evidence URLs to human "guardians" who fact check misinformation. Lim et al. [2017] created a framework in which users could provide feedback on whether evidence produced by an automatic system was relevant. Methods such as these show that the solution to these problems does not have to completely remove humans from the loop. Instead, they can supplement the superior fact-checking skills of human experts by speeding up their workflow.

Another interesting technological approach is the idea of "inoculating" people against false information. The idea is that if people are provided with examples of disinformation, they will know how to avoid falling for it. Roozenbeek and van der Linden [2019] created a game[34] which casts users as fake news producers who have to learn how to use common disinformation strategies to deceive others. Their findings suggested that after playing the game, users were more psychologically resistant to false information.

[29] https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

[30] https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html

[31] https://facebook.com/formedia/blog/working-to-stop-misinformation-and-false-news

[32] https://theguardian.com/technology/2021/feb/08/facebook-bans-vaccine-misinformation

[33] https://www.bbc.co.uk/news/technology-52903680

[34] https://www.getbadnews.com/

2.5.6 Related Areas

Clickbait

Tabloid Newspapers have always used attention-grabbing, often sensationalist headlines to draw shoppers into picking up a copy of the paper. This technique has also been adopted by many web-based news sites to get users to click on a story. Such news articles are commonly referred to as clickbait. There are both textual and non-textual methods for identifying clickbait, though a hybrid approach may often be the best [Chen et al., 2015]. Some methods have found great success by looking at headlines to identify clickbait articles [Anand et al., 2017], though there are certain issues with the datasets used; the real news headlines were all gathered from Wikinews, meaning they were from a range of authors on one site, whereas the clickbait headlines were gathered from a small handful of clickbait sites. Even assuming the clickbait sites only produce clickbait, it still has the problem that the system may be picking up differences in style between the sites more than between clickbait and news. Another popular technique is looking at the stance of the body with respect to the headline [Bourgonje et al., 2017]. Clickbait articles tend to have bodies that do not particularly support the headline.

While clickbait does not necessarily need to be fake, the techniques used have also been adopted by fake news websites. In cases where the deception is not politically or personally motivated, fake news articles are just a form of clickbait where the article has no basis in fact. So it makes sense for us to take in mind the things we learn from clickbait when looking at fake news.

Hyperpartisan News

Hyperpartisan news is news that presents a biased account from a certain side of the political spectrum. It has been well studied in recent years, as has risen in prominence since the 2016 US presidential election [Bhatt et al., 2018]. Fake and hyperpartisan news are highly related, as shown by Mourão and Robertson [2019], who suggest that fake news is more defined by partisanship than deception, and Recuero et al. [2020], who found connections between hyperpartisanship and disinformation when looking at the 2018 Brazilian presidential election.

Recently, the task of automatically identifying hyperpartisan news has been pursued.

Kiesel et al. [2019] ran a shared SemEval task, in which many teams created systems to predict if news articles were hyperpartisan. This was done using two datasets, one labelled manually per-article, and another much larger set labelled per-publisher. Teams achieved accuracies of over 80%, showing that the task was achievable. The most successful teams used word vectors as features, but manual features and even simple word n-grams still proved effective.

Potthast et al. [2018] looked at the stylistic features of hyperpartisan news, and successfully separated mainstream from hyperpartisan news. They also found that left and right leaning hyperpartisan news were more similar to one another than either was to mainstream news. Predicting media bias can either by done at article [Baly et al., 2020], or publisher [Baly et al., 2018] level, using both text features, and meta-features from outside sources. Predictions based on publisher have certain advantages, as one has more historical information to base a decision on, but it does rely on the dubious assumption that every article published by a specific news outlet is biased in the same way.

Irony and Humour

Figurative language is language where what is said is not literally what is meant. Humans understand the true meaning because they understand the context of what is being said. Computers, however, find the task of understanding the true meaning of figurative language very difficult [Reyes et al., 2012]. Humour and Irony are perfect examples of this. The reason humour is amusing is often because what is said has two meanings, one of which is funny. This idea underlies the semantic script theory of humour [Attardo and Raskin, 1991]. Irony is similar only in this case the opposite of what is said is usually meant [Wilson and Sperber, 1992]. It is almost like deception that the speaker wants you to see through.

This links to disinformation, because many fake news sites claim to be satirical. Satire is a form of humour that employs heavy use of irony to poke fun at current affairs and prominent figures. To understand satire, a reader needs to be aware of often quite elaborate context. Examples of satirical websites which mimic real news are `theonion.com` and `thedailymash.co.uk`. Given much fake news claims to be satire, looking for features of irony and satire in fake news may be helpful, even if it

is only to find instances where something claiming to be satire lacks the features of it. Past works have looked at using irony and humour features to identify satire and also compared satire to fake news [Rashkin et al., 2017, Horne and Adali, 2017]. Trying to understand context may help identify false information in the same way it helps with detecting irony.

Deception

There are two major types of deception: verbal and non-verbal. Non-verbal deception involves things that liars do which are physically observable, such as heart rate and eye movement. Verbal cues, on the other hand, are hidden within the content of lies. This often involves looking for linguistic clues or 'language-leakage', the linguistic features that liars struggle to mask when deceiving [Cody et al., 1984, Carlson et al., 2004].

Fake News is an interesting type of deception because the motivation is not always to deceive. While some authors of fake news might be trying to trick their reader, many are not. They might just be reporting a rumour that they have not scrutinised sufficiently, or they may have been told a lie and genuinely believe it. How do you tell when somebody is lying if even they do not know it? Another issue is that false information is often not 100% fake. There are often little seeds of truth within a good lie. The news writing style may also disrupt stylometric methods looking for author writing style.

Deception detection work has been done in other areas. Some works have looked at fraudulent dating profiles, finding that liars would often exaggerate [Toma and Hancock, 2010]. Work has been done on fake hotel reviews [Feng et al., 2012, Banerjee et al., 2015] where deep syntactic features and titles have been found useful. Others have looked for features of deception in fraudulent academic writing, which like our problem has the issue of style masking deception and of not everything being a lie [Markowitz and Hancock, 2014]. There are some examples such as satire where the author's intent is completely different, to amuse rather than decieve [Rubin et al., 2016]. By looking at many different domains we can get an idea of how the deceptive intent and level of belief in the non-truth affect the features of deception. It will also be interesting to see if there are any features that are universal across domains which may help future detection methods avoid making decisions based on features that are too genre specific.

We also need to consider how to explain these decisions to human readers. It is easy

to say that deceptive texts contain more negations, but that means almost nothing to the average person. An explanation needs to make sense. It would be more understandable to say that the article is more complex because these kinds of word are being used.

2.5.7 False Information in Online Communities

To understand false information, we must also understand the places and communities online where it is spread. Studies have been conducted looking at how fake news is spread on twitter, but also on websites like Reddit and 4chan [Zannettou et al., 2017, Hine et al., 2017]. Fake news is popular amongst hate communities as well as conspiracy theorists. What is shocking is that articles shared amongst these fringe groups often make their way to more mainstream social media such as Twitter and Facebook. It is important to think about who is writing these articles too. State actors have been accused of writing fake news to spread discord in other countries' democratic elections [ODNI, 2017]. The behaviour of such state actors has been studied by Zannettou et al. [2019]. Some fake news is also written by people with no political stake whatsoever[35], purely as a means of attracting traffic and therefore advertising revenue to their websites. Understanding how these authors use language when they do not believe or even care about what they are writing may provide insights applicable to fake news. It is important to be aware of where false articles originate, both so we can understand the audience of fake news and also so we know where to look for data when building datasets.

Various works have looked at the way disinformation is spread online over time. For example, Shao et al. [2016] created a platform, Hoaxy, which investigated the way in which false information was spread across the Twitter platform. Del Vicario et al. [2016] compared user reactions to conspiracy theories and science news, finding that consumers of both types of news have similar consumption patterns, but exhibit different cascades. Other works have looked at the way that Rumours spread across social media. A survey by Zubiaga et al. [2018] details research into the areas of rumour detection and resolution. We have already discussed some of these approaches above in section 2.5.4.

Health Communication is an area in which the investigation of disinformation propagation is of paramount importance. Some works have looked at the tactics used by the anti-vaccination movement [Kata, 2010, 2012]. Broniatowski et al. [2018] looked at

[35] https://tinyurl.com/ybno4tzu

the activity of Twitter bots and Russian-affiliated trolls in the sharing of anti-vaccination messages. They found that many of the "content polluter" bots[36] spread anti-vaccination messages and that Russian Trolls amplified both sides of the debate to spread discord. In NLP, work has looked at building corpora relating to vaccination [Morante et al., 2020], and predicting the vaccine-stance of social media posts [Skeppstedt et al., 2017]. The Covid-19 pandemic has increased interest in such work, and new solutions are being produced for countering Covid-19 and vaccination misinformation [Li et al., 2020b, Medina Serrano et al., 2020].

Samory and Mitra [2018] looked at discussions on the subreddit `r/conspiracy` following dramatic events. They looked at three different groups of users: 'Veterans', who were already members of `r/conspiracy`; 'Converts', who were already users of Reddit but new to the subreddit; and 'Joiners', who are new to both Reddit and the subreddit. They found that dramatic events, such as shootings and disasters, could be located by looking for spikes in new membership. Joiners and Veterans became more involved throughout their tenure, the authors suggest that Joiners may be the Veterans of tomorrow. They did not specifically investigate the conversion process, but observed that Veterans posted more as soon as the event occurs, possibly recruiting Converts in the process. Some linguistic analysis was carried out; after events, engagement rose but comments became more confused and exhibited more confused argumentation.

2.5.8 Datasets

One problem with current work in false information is that there is a lot of variation in the data that is used, and a distinct lack of benchmark datasets. Even still, some datasets have been created with the intention of being used as standard. This section will describe some of the datasets that exist, as well as some of their common features.

The type of text contained within each dataset varies substantially. Different types of data will suit different tasks. Some datasets contain only short claims [Wang, 2017, Thorne et al., 2018a], which are mostly useful for fact-checking. Others contain social media posts, for example from Facebook [Tacchini et al., 2017, Santia and Williams, 2018] or Twitter [Mitra and Gilbert, 2015, Zubiaga et al., 2016]. These datasets can be useful for tasks such as rumour verification and network approaches to false information

[36] They define these as bots who "spread malware and unsolicited content".

detection. Finally, there are some datasets which contain full news articles [Horne and Gruppi, 2021, Shu et al., 2020, Popat et al., 2018b].

Another thing that varies between datasets is the source of the labels. Some work labels texts at a per-text level [Popat et al., 2018b]. For each text a label may be taken from a fact-checker, or even crowd-sourced. Other work labels text at a per-publisher level [Horne and Gruppi, 2021]. The advantage of such methods is that one can create much larger corpora. However, they rely on the assumption that a publisher of disinformation will exclusively release articles which constitute disinformation. This can be fine in some cases (e.g. a satire website like the Onion), but is a naive assumption in other cases (e.g. the Daily Mail).

One limitation of many false information datasets is that they are topically constrained. Many relate to politics, as this topic attracts a lot of fact checking, not to mention that the effects of false information in this area can be particularly harmful. Often systems are only evaluated on one dataset, or one type of data. For example, a study may focus on a specific event [Gupta et al., 2013]. This is not inherently a bad thing – as it is worthwhile studying such datasets, and the results may be applicable elsewhere – but it does mean that it is difficult to establish the generalisability of results. The media used is also often limited. Datasets are often taken from Twitter, or to a lesser extent Facebook or Sina Weibo, due to ease of data collection. Media such as WhatsApp, which has been prolific in spreading misinformation in India and Brazil [Reis et al., 2020], are often understudied, presumably because it is harder to gather private messages than public posts.

There are some alternative or related datasets that are available for comparison. Various works have created datasets of fake reviews [Ott et al., 2011], satirical articles [Rubin et al., 2016, Rashkin et al., 2017], and more niche social media platforms [Zannettou et al., 2017, 2018a, Papasavva et al., 2021]. By looking at data from different sources, and different communities of people, research may be able to answer more general questions about the characteristics of false information.

2.5.9 False Information Summary

So far, research has looked at the linguistic features of deceptive texts, including fake news. However, when looking at such texts, not enough attention has been paid to the

author's intent and whether or not they believe what they write. The distinction has been made in past work between unintentionally spread 'misinformation' and intentionally spread 'disinformation'. However, the linguistic similarities and differences between these two types of false information have not been deeply investigated; particularly in terms of belief and intent. More research is needed into specific areas of false information, such as conspiracy theories, medical rumours, and propaganda, as well as into the machine learning techniques necessary to facilitate the work.

2.6 Literature Review Summary

In this chapter, we have provided a background in the key areas of research related to this thesis. We have introduced Natural Language Processing and Corpus Linguistics, and described key methods and concepts from these areas. Various works have been highlighted looking at language change and the study of online communities. Finally, we introduced the concept of False Information, and described current approaches to its study and detection.

Future chapters will expand on some of the work described in this chapter. In particular, we will look at previously unstudied sources of false information, namely April Fools articles (Chapter 3) and Flat Earth fora (Chapter 6). By studying new forms of false information, we will gain an understanding of the generalisability of the features of false information.

We are also interested in the analysis of sub-groups within communities. In Chapters 4 and 5, we will adapt existing language change methods for looking at sub-groups. These methods will then be applied to a Flat Earth community in Chapter 7, where we will try to better understand the types of user who make up conspiracy theory communities.

This analysis will involve bringing together many of the methods we discussed from NLP and Corpus Linguistics. Through the application of various techniques, we aim to learn more about the language of false information and answer the research questions outlined in Chapter 1.

Chapter 3

Linguistic Analysis of False Information

3.1 Introduction

As mentioned in Section 2.5, there are various different approaches to studying and detecting false information. By looking at different aspects, we can better understand the phenomenon in general. This thesis focuses on the linguistic dimension of false information, and this chapter, in particular, will try to identify linguistic features of disinformation by looking at a case study of April Fools hoax news articles. Using a combination of methods from corpus linguistics and NLP, we gain an insight into the way language is used in April Fools hoaxes compared to both genuine and "fake" news. The insight we gain will increase our understanding of the features of false information, and thus contribute to answering RQ1 from Section 1.3.

3.1.1 April Fools: An Interesting Case Study of Disinformation

People celebrate April Fools day each year on April 1st by playing pranks on each other for hilarity's sake. This tradition has transferred over to the traditional media, the most famous example of which is the BBC's 1957 'Swiss Spaghetti Harvest' film[1], which tricked many UK television viewers into believing that a farm in Switzerland grew spaghetti as a crop. With the rise of the web, news sites and companies started

[1] http://news.bbc.co.uk/onthisday/hi/dates/stories/april/1/newsid_2819000/2819261.stm

releasing annual hoaxes.

One of the main differences between April Fools articles and typical deceptive texts is the author's intent. The author of an April Fool is not trying to deceive so much as amuse. In this way, April Fools hoaxes are similar to Irony and Satire, which expect the reader to understand based on context that what is literally being said is not true. April Fools are a type of False Information where we know the intent of the author, which makes them an interesting dataset. The purpose of this chapter is not to isolate the features of intent, but rather to introduce a dataset of known deceptive intent as a comparison to more traditional "Fake News" datasets. Future work should look at the complex problem of isolating intent, and could use this dataset. However, we deem this beyond the scope of this chapter and thesis.

By using April Fools news hoaxes, we can look at a dataset of verifiable false bodies of text spanning back 14 years. Similar work with satirical news articles has yielded interesting results, such as that of Rubin et al. [2016], who trained an SVM classifier to predict satirical news with 90% precision. While it is true April Fools hoaxes are not completely similar to 'Fake News', mainly in terms of motivation, our hypothesis is that they will provide insight into the linguistic features put on display when an author is writing something fictitious as if it is factual.

Throughout this thesis, we will argue that looking at different types of false information, as well as examples from different origins, is crucial to gaining a full understanding of language in these texts. This chapter offers one such example, highlighting a source of disinformation that has so far been unstudied, but which has attractive features that lend itself to analysis. April Fools hoaxes may not be a "dangerous" form of false information, but they provide an unambiguous testing ground from which we may learn more about the general features of disinformation.

The main contributions of this chapter are:

- Introducing a new dataset of hoax April Fools articles, providing a novel False Information dataset with known deceptive intent.
- Investigating the linguistic features of April Fools hoaxes, particularly how they relate to features of deception and humour.
- Discussing how these features may be useful in the detection of Fake News.

3.2 Background

As April Fools reside in a space somewhere between deception and humour, we will provide a brief background in the areas of deception detection and humour recognition. We will also discuss current NLP approaches to Satire and 'Fake News' detection. These topics have already been touched upon in Section 2.5, but this section will go into more detail, particularly into important linguistic features.

3.2.1 Deception Detection

Deception research often focuses on 'non-verbal' cues to deception, e.g. eye movement. However, we are interested in the verbal cues to deception, i.e. the features hidden within the text. Without non-verbal cues, humans identify deception with very low degrees of success [Masip et al., 2012]. Much of the research on verbal cues of deception has been completed in the context of Computer Mediated Communications (CMC). This type of communication can be either spontaneous (synchronous) or preplanned structured prose (asynchronous), such as news, which is of interest for the present research.

Works in synchronous deception detection have involved looking at text from spoken and written answers to questions [Newman et al., 2003], email [Keila and Skillicorn, 2005], and chat-based communication [Hancock et al., 2007, Derrick et al., 2013]. Carlson et al. [2004] provide a good overview of how different factors can affect the deception model, such as medium, the liar's social ability, and the author's motivation. There are certain groups of features that these works suggest are present in deception. One of these groups is 'Cognition Features'. Lying requires a higher level of cognition than telling the truth so often lies seem to be less complicated and more vague. There is also a tendency towards negative emotional language because liars feel guilty about lying. Certain features suggest that liars have more distance from the story they are telling, e.g. reduced pronouns and details.

These works are useful for looking at the linguistic behaviour of liars, but they do not carry over too well to asynchronous communication, where a deceiver can edit and revise what they have written. Toma and Hancock [2010], looking at fake online dating profiles, found that certain features of synchronous deception were not present in asynchronous deception. Liars also more frequently exhibited exaggeration of their

characteristics. Other works have looked at fake hotel reviews [Ott et al., 2011, Banerjee et al., 2015]. Features relating to understandability, level of details, writing style, and cognition indicators provided useful clues for identifying fake reviews, though some features may have been genre-dependent. Markowitz and Hancock [2014] looked at fraudulent academic writing, an area similar to the news domain where a formalised writing style may mask certain stylistic features of deception. Fake works exhibited overuse of scientific genre words, as well as less certainty and more exaggeration. Stylometric approaches to looking at deception have included Afroz et al. [2012] who found that certain style features seemed to leak out even when an author was trying to hide them or imitate someone else. In some ways April Fools articles are an example of imitation, in which an author is writing a fictional article, mimicking the style of real news.

3.2.2 Fake News

Conroy et al. [2015] provide an overview of computational methods that can be used to tackle the problem of Fake News, including linguistic approaches. Oshikawa et al. [2020] carried out a more recent survey of NLP approaches to Fake News detection, showing various formulations of the task and state of the art results. This chapter is most interested in linguistic approaches which highlight important features for distinguishing fake from genuine news. Current linguistic research into detecting fake news includes Pérez-Rosas et al. [2018] who used features from LIWC [Pennebaker et al., 2001] for the detection fake news. They found that genuine news contained more function words and negations as well as more words associated with insight, differentiation and relativity. Fake News, meanwhile, expressed more certainty and positive language, and focused on present and future actions. These results are interesting, but it must be considered that the dataset used was crowdsourced using Amazon Mechanical Turk, meaning the authors of this news were unlikely to be accustomed to writing news articles. Horne and Adali [2017] found fake news to be a lot more similar to satire than normal news and also that the title structure and use of proper nouns were very useful for detecting it. Rashkin et al. [2017] found that features relating to uncertainty and vagueness are also useful for determining a text's veracity.

3.2.3 Humour Recognition

Unlike most deceptive texts, April Fools articles have a motivation of humour. Bringing ideas in from the area of humour recognition therefore may help us characterise hoax articles. Much of the work in humour recognition has focused on detecting humour in shorter texts such as one-liner jokes.

Mihalcea and Strapparava [2005] showed that classification techniques can be used to distinguish between humorous and non-humorous texts. They used features such as alliteration, antonymy, and adult slang in conjunction with content features (bag-of-words). Mihalcea and Pulman [2007] discussed the significance of 'human-centeredness' and negative polarity in humorous texts. Reyes et al. [2009] looked at a corpus of one-liners and discussed their features. Reyes et al. [2012] investigated the features of humour and contrasted to those of irony.

3.2.4 Irony

Irony is a particular type of figurative language in which the meaning is often the opposite of what is literally said and is not always evident without context or existing knowledge. Wallace [2015] suggests that to create a good system for irony detection, one cannot rely on lexical features such as Bag of Words, and one must consider also semantic features of the text. Reyes et al. [2012] created a dataset generated by searching for user-created tags and attempted to identify humour and irony. The features used to detect irony were polarity, unexpectedness, and emotional scenarios. Van Hee et al. [2016b] investigated annotated ironic tweet corpora and suggested that looking at contrasting evaluations within tweets could be useful for detecting irony. Van Hee et al. [2016a] also created a system to detect ironic tweets, looking beyond text-based features, using a feature set made up of lexical, syntactic, sentiment, and semantic features.

3.2.5 Satire

Satire is a form of humour which pokes fun at society and current affairs, often trying to bring something to account or criticise it. This is often achieved using irony and non-sequitur. Satire is similar to April Fools in that the articles are both humorous and often

untrue. One difference is that satire often tends to be political, whereas April Fools are usually more whimsical and varied.

Burfoot and Baldwin [2009] created a system to identify newswire articles as true or satirical. They looked at bag-of-words features combined with lexical features and 'semantic validity'. Rubin et al. [2016] used linguistic features of satire to build an automatic classifier for satirical news stories. Their model performed well with an F1-Score of 87% using a feature set combining absurdity, grammar, and punctuation.

3.3 Hoax Feature Set

The purpose of this chapter is to identify the features of April Fools articles, and to see if what we learn is also true of fake news, and possibly disinformation more generally. We want to avoid highly data-driven methods such as bag-of-words because these will learn content and topic-based features of our specific dataset meaning we would not necessarily learn anything about April Fools or deception more generally. We specifically look at the use of features from the areas of deception detection and humour recognition. A set of features used in past literature were selected, and together form our hoax feature set.

Some previous works have used LIWC [Pennebaker et al., 2001] to capture Neurolinguistic features of deceptive texts. While we did not use LIWC directly, we did consider important LIWC features from previous work when devising our own features.

For many of our features, we utilise tokenisation and annotations from the CLAWS Part-of-Speech (PoS) tagger [Garside, 1987] and the UCREL Semantic Annotation System (USAS) [Rayson et al., 2004]. The code we used for extracting features, including the output from CLAWS and USAS, are available for reproducibility purposes with the rest of our code[2].

The features we used have been split into seven categories so as to logically group them together to aid analysis and understanding of the results. These categories are: Vagueness, Detail, Imaginative Writing, Deception, Humour, Complexity, and Formality. By splitting our feature set into subsets, we can better understand how different types of feature behave on a higher level than looking at individual features.

[2] `https://github.com/dearden/april-fools`

For example, we may find that certain subsets of features are more useful than others. All features were normalised between 0 and 1.

Vagueness features aim to capture the idea that hoax articles may be less detailed and more ambiguous because the stories are fabricated. Ambiguity was captured by calculating the proportion of words in a text for which there were multiple candidates for annotation. Three types of ambiguity were used: Part-of-Speech Ambiguity, Semantic Ambiguity, and WordNet Synset Ambiguity.

Vague descriptions might use more comparative and superlative words as opposed to hard, factual statements [Ott et al., 2011]. Groups of PoS and Semantic tags were gathered to represent exaggeration, degree, comparative, and superlative words.

Detail features are almost the opposite of vagueness. Genuine news articles should contain more details because the events described actually happened. Increased cognition is needed to invent names and places in a text. For this reason we look at the number of proper nouns in a text. Similarly, a fake article may avoid establishing minute details such as dates. We therefore look at Dates, numbers, and Time-related words. Motion words, spatial words, and sense words also establish details that may be less present in deceptive texts.

Imagination features have been used in deception research by Ott et al. [2011], based on the work of Rayson et al. [2002], which involved comparing informative to imaginative texts. It is worth noting that we are comparing informative texts to pseudo-informative texts, rather than informative to openly imaginative texts. However, they were previously useful in detecting deceptive opinion spam [Ott et al., 2011], so we evaluate their use here. Rayson et al. [2002] identify different PoS tags that are more present in imaginative and informative writing. We used tags that were highlighted from the following PoS groups: conjunctions, verbs, prepositions, articles, determiners, and adjectives.

Deception features are the features of synchronous verbal deception. We include them to investigate if any of the features of spontaneous deception are preserved in spite of a change in medium. Features of asynchronous deception are more relevant to this task and have been distributed between more specific categories, such as Complexity and Details. These synchronous deception features are: First-person pronouns, Negative Emotional Language, and Negations.

Humour features are those from the area of humour recognition. As with deception, some humour features (notably ambiguity) fit better into other categories. The humour features used were: Positive emotion, Relationships, Contextual Imbalance, Alliteration, and Profanity. Contextual Imbalance is characterised as being the average similarity of all adjacent content words in the text. Similarity was calculated by comparing the vectors of words using the in-built similarity function of spaCy[3]. Positive Emotions and Relationships were both gathered using USAS semantic categories. Profanity was gathered from a list of profanities banned by Google [4]. Alliteration was measured by calculating the proportion of bigrams in the text that began with the same letter.

Formality features aim to capture elements of style in news documents that may show how formal they are. April Fools may be generally less formal or have less editorial oversight. We used three features based on aspects of the Associated Press (AP) style book[5]: AP Number, AP Date, and AP Title Features. These features checked if the text obeyed AP guidelines in their writing of numbers, dates, and titles. An example of an AP guideline is that all numbers under 10 must be spelled out (e.g. 'four' as opposed to '4'). Spelling mistakes were also counted and used as a feature, using the enchant spell checker[6].

Complexity features represent the structure and complexity of an article. They comprise: punctuation, reading difficulty, lexical diversity, lexical density, average sentence length, and proportion of function words. Punctuation was the number of punctuation marks in the text, found using a regular expression. To calculate the reading difficulty feature, we used the Flesch Reading Ease index [Flesch, 1948]. This method uses the number of words per sentence, and syllables per word to calculate a score. The higher this score, the more readable, though for our features we inverted the value so a higher value meant greater complexity. Flesch Reading Ease has been criticised for being crude and out-dated [Hartley, 2016], but for our purposes, it is still useful as a simple indication of complexity. For function words, we used a list from Narayanan et al. [2012].

[3] https://spacy.io/

[4] https://github.com/RobertJGabriel/Google-profanity-words/blob/master/list.txt

[5] https://www.apstylebook.com/

[6] https://github.com/rfk/pyenchant

3.4 Data Collection

3.4.1 April Fools Corpus

When building our dataset, the first challenge we faced was finding news articles that were definitely April Fools hoaxes. One cannot simply collect all news articles from April 1st as the majority of news from this date is still genuine, and it is infeasible to manually go through all news published on this day every year. So instead we used a website that archives April Fools each year[7]. The links published on this site are crowd-sourced so there are some issues arising from the fact that only the popular/amusing hoaxes are uploaded. However, this problem is fairly minor; in fact crowd sourcing may serve to diversify the kinds of website from which hoaxes are sampled. The site archives April Fools articles from 2004 onwards, providing 14 years of hoaxes[8].

We used Beautiful Soup [Nair, 2014] to scrape all of the hoax links. We performed preprocessing to remove hoaxes that one could tell did not constitute a news story from the URL. Next we processed the linked webpages, extracting the headline and body of each hoax separately. The wide range of sites in the corpus made automatic scraping too error-prone, so the final approach was largely manual. Efforts were made to ensure no boilerplate or artefacts from the website were included as these could have caused the classifier to pick up features such as the date as being features of April Fools. For the same reason, we also removed any edits to the article disclosing its April Fools nature.

There were various categories of April Fools articles found, the most common of which were news stories and press releases. News stories are distinct from press releases which we classed as texts that are self referential; usually taking the form of announcements or product reveals. For example, a press release might be a website announcing that they have been bought out by Google, whereas a news story might be an article by the BBC saying that Google has bought out said company. Press releases were manually filtered out for the present study in order to keep the focus on news, and to avoid the features of press releases obscuring those of April Fools articles. This resulted in a final April Fools (AF) corpus[9] comprising of 519 unique texts, spread across 371 websites.

[7] `https://aprilfoolsdayontheweb.com`
[8] All articles were collected in 2018.
[9] `https://doi.org/10.17635/lancaster/researchdata/512`

3.4.2 News Corpus

To create a comparable corpus of genuine news articles, Google News was utilised to automatically scrape news articles from the 4th–5th April of the same years (2004–2018) This time range was chosen so the kinds of topics in the news would be of a similar nature. We will refer to these articles as "NAF" articles. The stories were found using 6 search terms that aimed to catch similar topics to those represented in the AF articles. We did this to avoid learning about the differences in topics of articles rather than whether or not an article is a hoax. The following search terms were selected based on a manual analysis of the topics covered in the AF corpus: "news", "US", "sport", "technology", "entertainment", and "politics". Despite our efforts, it was difficult to match the topic distribution exactly: not all the websites in the AF corpus have archived articles going back to 2004. We acknowledge this is a problem but do not consider it too critical as the features we are looking for are not data-driven and so should not be influenced by topic. We then took all of these URLs and automatically scraped the text using the newspaper python package[10]. Using this method we scraped 2,715 news articles.

For each year (2004–2018), we selected the same number of articles as there were in the AF corpus. The 519 AF articles were spread over 371 websites, the most common of which occurred 19 times. To try and match this distribution, we capped the number of articles that could be taken from any given site at 20. Once we had selected our genuine (NAF) articles, we manually checked the text of each article to ensure that the full text was scraped correctly and that the text only contained the news article itself, without boilerplate noise. We went through the same process as for the AF articles of removing any texts that did not fit in the category of News, such as personal blogs. When an article was removed, we replaced it by choosing a news article from a later page of the Google search that found it. Once this process was finished, we had an NAF corpus of 519 articles spread over 240 websites. Table 3.1 shows a summary of the corpus, which is made available for further research. April Fools articles contain fewer words on average, which will be accounted for by normalising frequency features by document length. Both AF and NAF articles vary significantly from the mean in their lengths.

[10] `http://newspaper.readthedocs.io/en/latest/`

Table 3.1: Summary of April Fools (AF) and Non-April Fools (NAF) corpora.

	Articles	Websites	Avg Words	Std Words
AF	519	371	411.9	326.9
NAF	519	240	664.6	633.2

3.4.3 Limitations

This is a small dataset and has various notable limitations. The genuine articles tend to be from a smaller pool of more established websites as it is these websites that are more prominent when searching for news online. Only news articles are contained in the dataset, further work may extend to blogs and press releases. Sometimes the distinction between blogs and news is arbitrary but we tried to be consistent. Multiple genuine news articles occasionally cover the same story, but this was rare and no one story was ever repeated more than twice. A minimum length of 100 characters was enforced to remove anomalous texts such as video descriptions, however this may have removed some genuine articles. While we bear them in mind, we do not see these limitations as major barriers to the research. We will analyse the data using both quantitative and qualitative techniques that allow us to take a deep dive into the data and understand the language being in April Fools articles for the first time. We do not believe a significantly larger corpus could be built in a reasonable time period.

3.5 Analysis

3.5.1 Classifying April Fools

To evaluate the comparative strength of our feature groups for predicting hoaxes, we used a Logistic Regression classifier with 10 fold cross-validation. We used default parameters of Logistic Regression (from scikit-learn[11]), with standardization to zero-mean and scaling to unit variance ($x' = x - \bar{x}/\sigma$). A basic Logistic Regression classifier serves our needs as we are primarily concerned with investigating the behaviour of features with an interpretable model, and not maximising classification accuracy

[11] `https://scikit-learn.org`

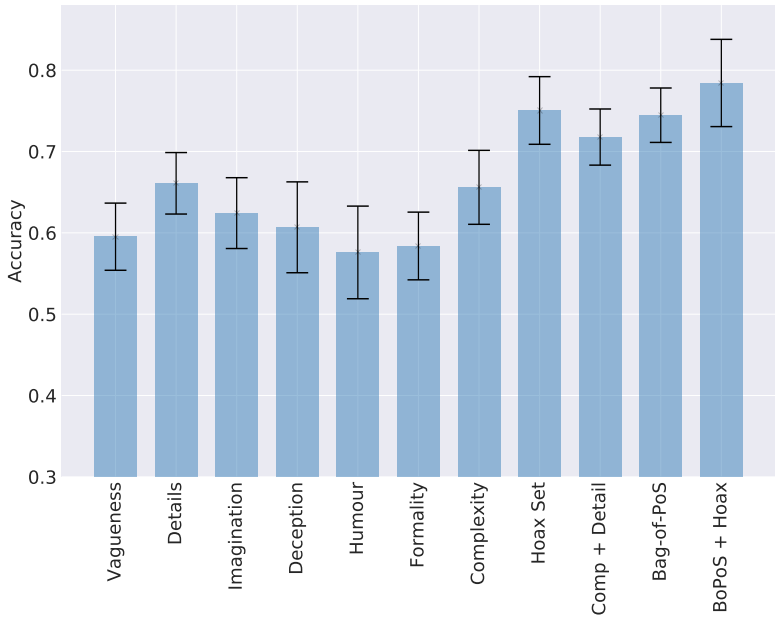


Figure 3.1: Mean accuracies of Logistic Regression classifiers across 10 Fold Cross-Validation. Error bars show standard deviation of accuracies across the 10 folds.

through tuning or more elaborate classifiers. The results of these classifications can be seen in Figure 3.1.

From the classification results, we can see that our features provide some information to differentiate between April Fools hoaxes and genuine news articles. The results are not as high as the F_1-Score of 0.87 found by Rubin et al. [2016] for the related task of satire detection, though they are similar to results from fake news detection, such as those of Horne and Adali [2017] who achieved an accuracy of 71% using bodies of text and 78% using headlines.

Looking at the individual feature groups, Complexity and Detail Features perform best, though not as well as the full Hoax Set. Deception literature suggests that deceptive accounts contain fewer specific details and are generally less complex [Carlson et al., 2004]. Humour performing badly is not surprising as understanding the joke of an AF hoax requires a lot of context and pre-existing knowledge. The features of humour we used in the Humour feature-set were relatively simplistic; more complex, context-aware features may be needed to identify the humour in April Fools hoaxes. The poor performance of Formality features could suggest that AF Hoaxes are still written to the same journalistic guidelines and standards as their genuine counterparts.

Given the success of the Complexity and Detail features, we classified articles using only these features, achieving an accuracy of 0.718, not far from that of the entire Hoax

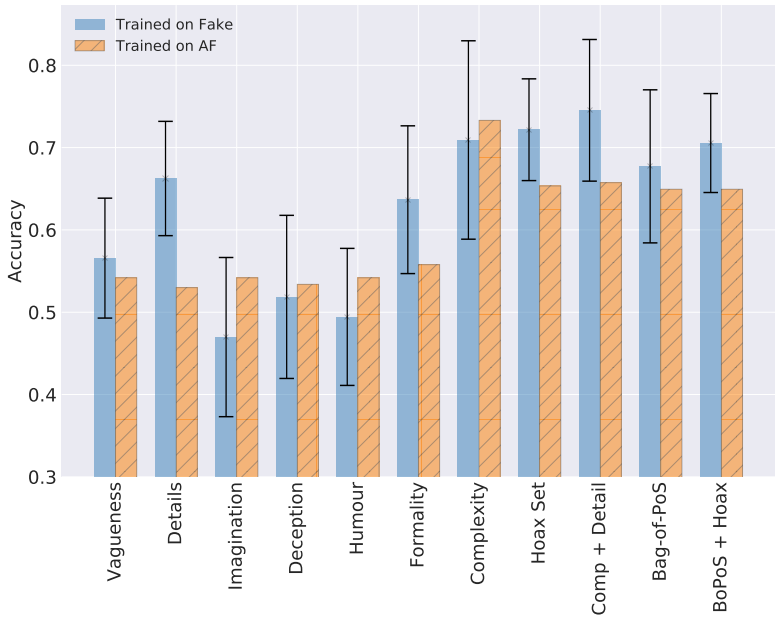


Figure 3.2: Accuracies of Logistic Regression classifiers for detecting fake news, trained on Fake News using 10 fold cross-validation and April Fools. Error bars show standard deviation of accuracies across the 10 folds.

Set (0.750). This further suggests that looking at details and complexities within a text are crucial when trying to determine if an article is a hoax.

We looked at a non-tailored Bag-of-Part-of-Speech (BoPoS) approach, to compare our curated features to a more data-driven approach. Each PoS tag in CLAWS is used as an individual feature, the occurrences of which are counted for each document. BoPoS was chosen over the more standard Bag-of-Words (BoW) approach because BoW is prone to identifying differences in content and topic, rather than style. BoPoS achieved an accuracy of 0.745, similar to the hoax set. This is not overly surprising as many of the hoax features were part-of-speech counts. These sets do not completely overlap, however. When the hoax set was added to the BoPoS features, the classifier improved its accuracy. This suggests that the non-part-of-speech features in the hoax set provide useful additional information. BoPoS performing similarly to the hoax set therefore suggests that there must be additional PoS frequencies which characterise AF hoaxes.

3.5.2 Classifying "Fake News"

Next, we see if we can use the same feature set to effectively identify Fake News. For this we used the fake news dataset introduced by Horne and Adali [2017]. This dataset consists of a mixture of articles gathered from well-known fake news outlets and

legitimate sites as well as articles gathered by Buzzfeed for an article about fake news in the 2017 election[12]. This is a small dataset (250 articles) split evenly between real and fake. The classification results, again using logistic regression and cross-validation, for fake news can be seen in Figure 3.2. For each feature set, one classifier was trained on fake news and evaluated using 10 fold cross-validation, and another trained on the entire April Fools corpus and tested on the entire fake news corpus.

The classifier trained on fake news using the hoax features achieved an accuracy of 0.722, similar to that achieved by the classifier trained on the hoax features for April Fools (0.750). This suggests that at least some of the features useful for detecting April Fools hoaxes are also useful in the identification of deceptive news. Complexity features performed well on the fake news dataset (0.709), performing almost as well as the full Hoax Set. Details were useful, as before, but Vagueness features performed substantially less well.

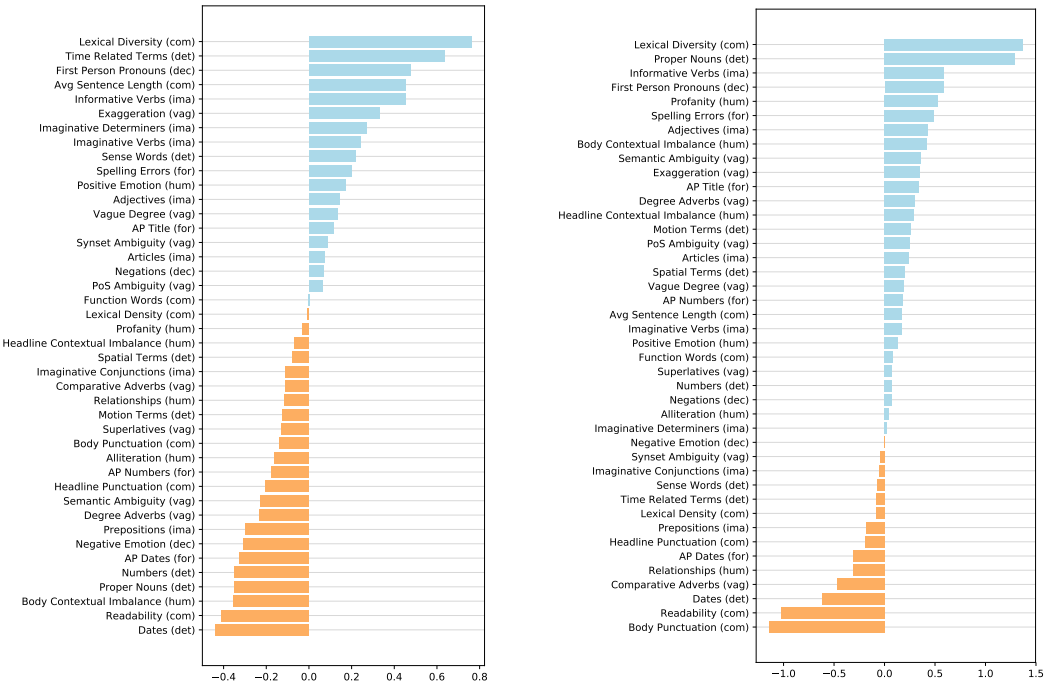
When trained on April Fools and predicting fake news with the Hoax Set, an accuracy of 0.653 was achieved. It is possible that some of the same features are useful but their behaviour is different for fake news. Still, the accuracy is not far off the Hoax Set, so there may be some features that manifest themselves similarly for both AF hoaxes and fake news. Finding these features could provide insight into deception and disinformation more generally.

BoPoS performed less well on Fake News, with an accuracy of 0.677, suggesting that PoS tags are not as important when looking at fake news. This, combined with the fact that the hoax features maintained a similar accuracy and complexity features did almost as well as the entire feature set, suggests that the structural features are more important when identifying fake news. BoPoS also did worse when trained on AF and tested on fake, and its drop in accuracy was similar to that of the hoax set. This suggests that there are some PoS tags that are distributed similarly for April Fools and fake news.

3.5.3 Individual Feature Performances

To see how important individual features were to the classifier, we looked at Logistic Regression weights as shown in Figure 3.3. For some features it is interesting to see how they are distributed. To this end, frequency density plots for some of our features

[12] `https://tinyurl.com/jlnd3yb`



(a) April Fools

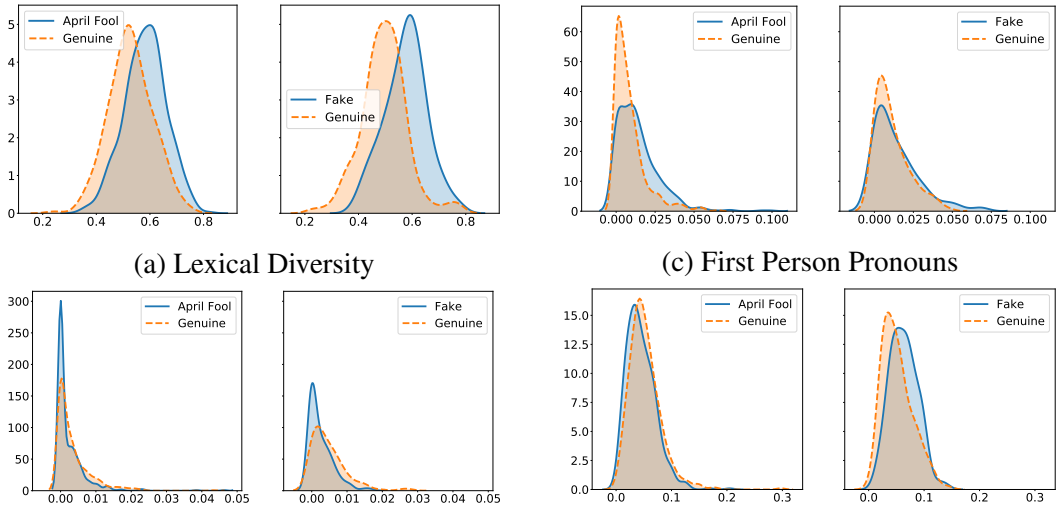
(b) Fake News

Figure 3.3: Logistic Regression weights for the Hoax Set. A large positive weight suggests an important feature of April Fools / Fake News and a large negative weight suggests an important feature of genuine news.

are provided in Figure 3.4. We tested the significance of differences in features between corpora using a Mann-Whitney U Test [Mann and Whitney, 1947]. Unless otherwise stated, the differences we discuss are statistically significant ($P < 0.025$).

There are differences in structural complexity between AF and genuine articles. Lexical Diversity is the most highly weighted feature. As we can see in Figure 3.4a, the feature separates hoaxes from genuine articles quite significantly. This could mean hoax texts use more unique words, but it could also be down to the difference in length. High values of lexical diversity correlate to shorter texts and, as we can see in Table 3.1, the AF articles are shorter, on average. This does still show, however, a difference in complexity. Average sentence length and readability being important features also suggests a difference in complexity. Genuine articles slightly tend towards a shorter average sentence length. NAF articles also tend towards being slightly more difficult to read, though again the difference is not huge.

The story is similar with Fake News – with structural complexity providing key features. Lexical Diversity behaves the same as in April Fools (Fig 3.4a). This could again be something to do with average document length, but also could suggest a higher



(a) Lexical Diversity

(c) First Person Pronouns

(b) Dates

(d) Proper Nouns

Figure 3.4: Density plots of notable features.

proportion of unique words. Reading difficulty also remains important, though the difference in distribution between fake and real is far more prominent, with genuine articles generally more difficult to read. This means that they generally contain longer sentences and words with more syllables. This difference could suggest that fake news articles are more simplistic than genuine texts. Body punctuation, a feature not weighted as highly for AF, appears to be very important for identifying fake news. More punctuation implies complex structures, such as clauses and quotes.

There are also differences in the level of detail between AF and NAF. Genuine articles tend to contain fewer time-related terms. This seems to go against the idea that genuine articles contain more detail. However, if you look at the occurrences of this feature in the text, the most frequent time-based term is 'will'. This combined with the fact that April Fools tended towards fewer dates (Fig 3.4b) and numbers suggests that AF hoaxes refer to events that will happen, but do so in vague terms. This backs up the idea that April Fools are less detailed and more vague. There are also more references to the present. AF hoaxes seem to be more interested in the present and future than the past. AF hoaxes containing fewer dates is interesting, as one might expect that an AF article would mention the date more than a regular article. This is true as far as references to April are concerned, April Fools had more of those. However, the number of references to the month was roughly the same (April Fools actually had slightly fewer overall), though for genuine news it was spread across more months. This may be because real news stories are the culmination of multiple past events that need to be referenced in the

story. More significant than references to months, were references to days of the week. Genuine articles contained many more, which backs up the idea of real texts building more detailed stories.

The distribution of proper nouns between AF and NAF is fairly similar (Fig 3.4d), possibly skewing towards fewer in April Fools. This could suggest fewer details, i.e. names and places, being established in the AF articles. Similarly to complexity, the differences in details are not huge, but do seem to be present.

The detail features do not quite behave the same in Fake News articles as in AF. Proper nouns are one of the most important features for characterising fake articles (Fig 3.4d). However, unlike AF hoaxes, fake articles tend towards containing more proper nouns than genuine articles. This does not suggest less detail. When looking at the corpus, Fake News articles seem to use a lot of names, often the same ones, such as "Trump" and "Hillary". Interestingly, they massively over use the name "Hillary", both suggesting that they are less formal (using the lone forename of a politician), and also that they may have an obsession. Dates are the only other Detail feature to be weighted highly for fake news. Fig 3.4b shows that this feature behaves similarly as it did for AF hoaxes, though not as dramatically so. Fake articles are more likely to contain very few, or no, dates. These findings suggest that there are not the same types of difference in detail between AF and fake news, though detail does still hold some significance: it was the second best performing feature group.

Not all the important features link to detail and complexity. First person pronouns were an important feature for both AF hoaxes and fake news. The word 'we' was overused in particular by April Fools and to a lesser extent by fake news. This goes against the ideas from traditional deception detection [Carlson et al., 2004] that suggest liars use fewer first person pronouns. In our data, the fake texts use more self-references[13]. This could point towards false articles being more personal and less formal, rather than a feature of deception.

Some of the highly weighted features of fake news are not in common with April Fools. For example, profanity and spelling errors. Both could point towards a reduced level of formality. This would make sense as not being a feature of April Fools. AF writers are usually writing for outlets that publish genuine news, and so likely conform

[13]Though, for Fake News, this difference was not significant.

to many of the same standards as genuine news. Fake news, however, comes from less journalistically formal websites.

One of the most obvious differences between April Fools and Fake News in Figure 3.3 is that Fake News has a smaller group of features that are very important. Lexical diversity, proper nouns, body punctuation, and readability are significantly higher weighted than anything else. Three of these four features relate to structural complexity and the other to detail. This could suggest that, in the case of fake news, the 'fakeness' lies in the structure of the words rather than the words themselves.

Our results suggest that April Fools and Fake News articles share some similar features, mostly involving structural complexity. The level of detail of a document is also important for both AF hoaxes and fake news, though these features do not behave exactly the same way. Some of the features of deception are present in April Fools, notably those relating to complexity and detail but also first person pronouns, though their behaviour is reversed. The basic features of humour we gathered seem to be less important. A more advanced study of the humour would be required to identify it within the AF hoaxes. A successful approach would likely require substantial context and world knowledge.

To compare them to the findings from our feature set, and demonstrate how we can gain new insight by looking at features prominent in the data, as well as those from past literature, we looked at some of the PoS tags that were highly weighted by the BoPoS classifier. Some familiar features show up. Certain time-related tags such as 'quasi nominal adverbs of time' (e.g. "now", "tomorrow") and singular weekday nouns (e.g. "Monday") are highly weighted. Proper Nouns are also highly weighted for fake news in particular. Coordinating conjunctions (e.g. "and", "or") are a prominent feature of NAF articles. More coordinating conjunctions implies more detail and complexity. It is good to see that some of the most highly weighted parts of speech back up our finding that detail and complexity are important in defining April Fools articles and Fake News.

3.6 Corpus Linguistic Analysis

In Section 3.5, we performed a comparison of AF hoaxes to genuine and fake news articles using NLP methods. Building on this, we will now perform a more qualitative,

corpus linguistic analysis to find examples that demonstrate the differences. To aid in linguistic analysis, we used the Wmatrix tool [Rayson, 2008] to analyse the April Fools articles.

Table 3.2 shows the five most key PoS tags in April Fools compared to genuine articles, sorted by log-likelihood (LL), a significance metric widely used in corpus linguistics. Table 3.3 shows a similar table, but with tags selected using Log-Ratio (LR), an effect size metric. These two metrics tell us slightly different things. LL tells us which words have the most significant difference between corpora, while LR tells us which have the largest difference, proportionally. Both these metrics were introduced in Section 2.2.

From these tables, we can learn more about the types of words that characterise April Fools hoaxes. The first main takeaway are that AF hoaxes refer more than genuine articles to future events. This makes sense, as AF articles often involve announcing something for the first time. It is also, presumably, easier to make something up if it has not happened yet. This preference towards the future may contribute to reduced detail in AF hoaxes. The second takeaway is that AF hoaxes refer directly to the reader. This ties into the idea from Section 3.5 that AF hoaxes are less formal than genuine news. The personal touch may be a by-product of the humorous element, or perhaps it makes readers more likely to be deceived.

Tables 3.4 and 3.5 show the equivalent tables for genuine news compared to AF. A notable difference here is that genuine articles use the past tense more than AF. This lines up with the fact that AF hoaxes were more focused on future tense. It seems that a key distinction is that genuine articles discuss events that have happened, rather than those that will.

The key PoS tags for genuine articles suggest that genuine news establishes details more than hoaxes. Using more 3rd person pronouns suggests that more people are mentioned in stories. The increased used of titles (e.g. “Miss”, “Dr”, “Lord”) could similarly suggest that more people are discussed, or it could suggest an increased formality. Past tense language could also link into detail, as more detail can be provided for past, rather than future, events. Coordinating conjunctions were used more in genuine news, which could suggest longer and more complex sentences. While this is not quite the same as detail, the concepts of complexity and detail go hand in hand –

Table 3.2: The five top features characterising AF articles, chosen using Log-Likelihood.

PoS Tag	**Part-of-Speech**	**Example**	**Log-Ratio**	**Log-Likelihood**
VM	Modal Auxiliary	*"The world of brewing **will** never be the same again"*	0.57	301.97
VBI	Be, infinitive	*"will **be** completed in 2016"*	0.75	225.77
PPY	2nd person personal pronoun	*"rolling in to a computer near **you**"*	0.88	190.25
PPIS2	1st person plural subjective personal pronoun	*"**We** thought **we** might share some..."*	0.80	178.28
VVI	infinitive	*"to once again **revolutionise**..."*	0.25	112.66

Table 3.3: The five top features characterising AF articles, chosen using Log-Ratio.

PoS Tag	**Part-of-Speech**	**Example**	**Log-Ratio**	**Log-Likelihood**
UH	Interjection.	*"**Oh**, and while you're here.", "**Hmm**..."*	0.99	47.73
VVGK	-ing participle catenative.	*"There are **going** to be 3D televisions."*	0.95	26.24
PPY	2nd person personal pronoun.	*"rolling in to a computer near **you**"*	0.88	190.25
PPIS2	1st person plural subjective personal pronoun.	*"**We** thought **we** might share some..."*	0.80	178.28
PPIO2	1st person plural objective personal pronoun.	*"Let **us** know if this news has you blue..."*	0.80	34.33

Table 3.4: The five top features characterising genuine articles, chosen using Log-Likelihood.

PoS Tag	Part-of-Speech	Example	Log-Ratio	Log-Likelihood
VVD	Past tense of lexical verb.	*"The company **said** the strong profit..."*	0.49	302.94
PPHS1	3rd person sing. Subjective personal pronoun	*"**he** said", "**she** told"*	0.83	275.66
VBDZ	Was	*"Pelosi's visit **was** criticised"*	0.63	149.30
CC	Coordinating Conjunction	*"and", "or"*	0.29	149.23
MC	Cardinal Number, neutral for number.	*"One", "2016", "6 Million"*	0.44	141.64

Table 3.5: The five top features characterising genuine articles, chosen using Log-Ratio.

PoS Tag	Part-of-Speech	Example	Log-Ratio	Log-Likelihood
PPHO1	3rd person sing. objective personal pronoun (him, her)	*"Rumors later placed **him** in New York"*	0.87	54.70
PPHS1	3rd person sing. subjective personal pronoun.	*"**he** said", "**she** told"*	0.83	275.66
VVD	past tense of lexical verb.	*"The company **said** the strong profit..."*	0.77	35.39
NNB	preceding noun of title.	*"**Lord** Heseltine", "**Miss** Miley Cyrus"*	0.64	60.62
VBDZ	was	*"Pelosi's visit **was** criticised"*	0.63	149.30

more complex sentences allow the conveyance of more detail.

In addition to PoS tags, we also looked for key semantic concepts, as represented by USAS tags [Rayson et al., 2004], for the AF and genuine articles. Generally, this was not particularly informative, as the key tags often corresponded to topics that were only in a single article, or a small pool of articles. It did, however, corroborate what we suggested based on parts-of-speech – that AF articles contain more of the future tense. We also saw some topical preferences that seemed to extend beyond random variations. AF hoaxes were more likely to use words relating to living creatures. Genuine articles, meanwhile, were far more prone to discussing heavy topics such as politics, war, and death. This suggests that AF articles tend towards more light-hearted topics.

We also looked at the key PoS tags for Fake News, compared to genuine and AF articles. These texts had some differences in common with AF hoaxes to genuine news. For example, fake news articles overused personal pronouns, though unlike with AF, future tense was not a significant feature. Fake news articles also had some common differences with genuine news to AF hoaxes. AF texts overused future tense compared to fake news, though not as substantially as with genuine. Fake news articles overused words such as “he” and “she”, similarly to genuine articles, which may suggest that this reporting of peoples’ actions is related to the news domain. Various detail-related features (such as coordinating conjunctions, plural proper nouns, and numbers) were overused by genuine news compared to fake news. These findings show that not all features of AF hoaxes apply to disinformation generally, but some features may be common, such as more casual, less formal language. It is worth bearing in mind that the fake news corpus is very small, and both topically and temporally limited, so this part of the analysis should be taken with a pinch of salt. However, it highlights the need to look at different types of disinformation to build a clearer picture of the general features of false information.

The results from the corpus linguistic analysis carried out in this section support the results from Section 3.5. They suggest that AF hoaxes are more casual and based around future events, while genuine articles contain more details and complexity. Fake news articles seemed to sit somewhere between genuine news and April Fools, sharing some features of both. Using corpus methods alongside NLP techniques helped to demystify the results of our classifier, helping to better explain its decisions and feature weights.

3.7 Conclusion

In this chapter, we have introduced a new corpus of April Fools hoax news articles. We also created a feature set based on past work in deception detection, humour recognition, and satire detection. Using this feature set, we built a system to classify news articles as either April-Fools hoaxes or genuine articles. The resulting accuracy of 0.750 suggests that the features we identified are useful in identifying April Fools hoaxes, though not without room for improvement. We then tested our system on a small dataset of fake news to see if April Fools hoaxes are similar enough to fake news that similar features can be used to detect both. An accuracy of 0.722 was achieved on the Fake News dataset, suggesting that these features are useful for both tasks.

We analysed our features using a combination of qualitative and quantitative techniques to observe the differences between April Fools hoaxes and genuine articles. This analysis suggests that the structural complexity and level of detail of a text are important in characterising April Fools. This was also the case for Fake News, though structural complexity seemed more important and the changes in details differed slightly from those in April Fools. Our findings suggest that there are certain features in common between different forms of disinformation and that by looking at multiple varieties, we can learn more about the language of disinformation in general. We also showed that by using a mixture of analysis techniques, we can gain far more insight than we can purely from classification. The corpus we have introduced will also be useful in wider fake news research by providing a dataset of news articles which are completely untrue, similar to how satirical news articles are already being used.

Despite similar features being effective at classifying both April Fools hoaxes and Fake News, we showed that not all these features behave the same way between the two text types. It is possible that some of these differences in feature behaviour come down to the deceptive intent of the texts. April Fools are an interesting form of disinformation because the author does not believe what they are writing and is not trying to deceive anybody. By looking at a wider variety of false texts, we can further understand the way that the author's motivation and belief affect the way false information is written.

This chapter has provided a new dataset for use in the area of fake news detection and has highlighted directions for future work, describing features useful for detecting April Fools articles and showing that they may also be present in fake news. The dataset,

and corresponding analysis, will contribute to our understanding of false information, in particular towards the identification (if they exist) of universal linguistic features of disinformation.

Chapter 4

Methods for Exploring Language Change of Groups

4.1 Introduction

In Chapter 3, we looked at the language of false information, and investigated the differences between April Fools and fake news. One limitation of this work, and many others in the literature, is that it treated language as static. Language is not static. The way people speak, and the meaning of the words they use, change constantly. These changes could be motivated by fashions, the introduction of new terms, or by real world events. When studying the language of false information in online communities, time is a crucial dimension that we cannot ignore. In Chapter 6 we will introduce a Flat Earth Reddit community which shut down due to an alleged influx of trolls. Cases like this show the need to look at how language evolves in false information communities. The aim of this chapter is to produce a toolbox of methods that can be used to observe the language change of groups within communities over short time-spans.

Despite language change being important, many of the traditional methods used in NLP treat language as static. The work that does address language change is largely focused on long-term change. Often, they aim to look at the change of language in general, for example by looking at the development of certain grammatical constructions [Gries and Hilpert, 2008], or the evolving meaning of words [Hamilton et al., 2016a]. This means that many of the methods have been designed and evaluated on corpora spanning decades, sometimes even centuries [Michel et al., 2011].

Our interest is in looking at the way language changes in communities, particularly online communities such as forums. Any language change in this setting will be comparably short term – certainly there has not been an abundance of data for long enough to have corpora spanning decades. Communities have varying levels of coherence, and often consist of changing sub-groups and membership. For example, within disinformation communities, you may find various types of user, e.g. believers, non-believers, and trolls. Developing methods to observe how these potentially competing language styles evolve over time would be a valuable contribution. We are interested in producing a set of methods to see whether language change manifests differently in different sub-groups of these communities, and how the language of such groups changes over short time-spans.

Research has looked at similar problems before, including online communities such as Reddit or forums. Several of these works look at language change. Examples of tasks performed in these works include looking at the way new terms develop online [Kershaw et al., 2017], and charting the linguistic similarity of users to communities over time [Danescu-Niculescu-Mizil et al., 2013]. The contribution of this work is the extension of this type of analysis to sub-groups, as well as the application of methods primarily used for looking at long-term change to this setting. All the methods in this chapter will aid us in answering RQ2 from Section 1.3, helping us understand how existing methods can be used and adapted to observe the language change of groups within communities.

In this chapter, we will demonstrate a range of language change methods from NLP and corpus linguistics on a corpus of UK House of Commons debates. Parliaments are an interesting example of a community, with relatively consistent membership, members who speak regularly, as well as known groupings and factions. Particularly of interest is how debates develop surrounding large events, such as the United Kingdom's exit from the European Union (Brexit). Any language change in this setting will be comparably short term – over a few years rather than decades or centuries – so any method to observe such change must be effective over a short timespan.

The language of political communities has been the focus of much research; including studies of polarization [Peterson and Spirling, 2018, Demszky et al., 2019], estimating political positions [Slapin and Proksch, 2008, Lauderdale and Herzog, 2016],

and looking at long term change in language usage by politicians [Jordan et al., 2019]. UK Political debate, specifically, has been frequently analysed in Linguistics [e.g. Wenzl, 2019], and occasionally in NLP [e.g. Abercrombie and Batista-Navarro, 2018].

Though parliamentary text is not false information, and nor is it an online community, the corpus is similarly structured to online forum data, which makes it an interesting case-study for testing methods that look at short-term language change in groups. Members of the community "post", or in parliament's case speak, regularly. Some users are members of the in-group, and some are not. The primary advantages over forum data, are that we can divide the user-base into known groups rather than having to guess, and the membership does not dramatically change over time. Another advantage is that we can sanity-check our methods by comparing their findings to our knowledge of the events that took place during the Brexit period. These characteristics make Hansard a preferable test bed for these methods over forum data. In Chapter 7, we will extend this analysis to online forums. This will not only help us understand how language has changed in the forums we will investigate, but it will also provide insight into how generalisable the methods described in this chapter really are.

The data and code from this chapter has been made publicly available for use in research. Data is stored in a database for easy access[1], and the code is found in two GitHub repositories: one for the methods[2], and another for the notebooks to run the experiments shown in this chapter[3].

4.2 Data Collection

4.2.1 The Data

For this work we are using data from the UK's parliamentary Hansard [4]. Hansard is a report of everything said in the houses of parliament. It describes itself as "Substantially Verbatim", meaning that it records all words spoken in parliament, albeit with repetitions and obvious mistakes removed. This makes it an interesting, though not entirely natural, corpus of spoken political language. Hansard covers both houses of

[1] https://doi.org/10.17635/lancaster/researchdata/514
[2] https://github.com/dearden/language_change_methods
[3] https://github.com/dearden/thesis_language_change
[4] https://hansard.parliament.uk

parliament: the Commons, where elected Members of Parliament (MPs) debate policy and laws, and the Lords, where Peers of the Realm can amend or reject bills passed to them from the House of Commons. For the purpose of this work, we are only going to use the Hansard records from the House of Commons, as this is the primary chamber of the UK's parliament.

As a brief primer for those not au fait with the UK political system, we will now give a brief overview. The House of Commons is the main UK legislative chamber, where MPs debate government policy and pass bills. There are two opposite rows of benches in this chamber: one set for the Government, consisting of the majority governing party (or coalition), and another for the opposition parties, the largest of which is called "the Opposition". Due to the UK's first past the post voting system, politics is dominated by two parties: the Conservative Party and the Labour Party. Members of Parliament are elected in General Elections which, since the Fixed Terms Parliament Act (2010), take place every 5 years (in theory more than in practice). A day in the House of Commons begins with a session of questions where MPs can question government ministers. This is followed by debate, where members can put forward motions and MPs take turns to debate them. At the end of a debate, there may be a vote (division).

In this chapter we use a dataset made up of House of Commons debates between the 2015 and 2019 UK General Elections (May 2015 – Dec. 2019). We chose this time range because we were interested in seeing if we could observe language change in groups based on the 2016 UK Referendum on membership of the European Union. This referendum had two sides: Leave and Remain. Leave won the referendum, but the majority of MPs supported remain prior to the vote. The key events of Brexit are described in Figure 4.1.

Though the UK actually left the EU on 31st January 2020, we chose to end the corpus with the December 2019 General Election. This was seen as the optimal cut-off point because in this election the Conservatives won by a significant majority, introducing a number of new MPs[5]. Adding new members so late in the corpus, with very little debate taking place between the election and the 31st seemed unnecessary, and potentially confusing. As it stands, the corpus contains the complete proceedings of the two parliaments, formed in the 2015 and 2017 general elections. This time range

[5] https://www.politicshome.com/thehouse/article/class-of-2019-meet-the-new-mps

Figure 4.1: A timeline of Brexit, highlighting key events within the time range of our corpus. Events were selected from Walker [2021]’s timeline of Brexit. Some events have been merged together if they took place within a short span of time. For this reason, we have not provided exact dates on this figure.

also covers three Conservative Prime Ministers: David Cameron (2010-2016), Theresa May (2016-2019), and Boris Johnson (2019-present). There is no specific endpoint of Brexit debate in the House of Commons, and in future the timeframe of the corpus could be expanded in either direction, but because of its proximity to the UK leaving the EU, and the relatively clean nature of ending before a new election, the 2019 election serves as the best possible point to end the corpus.

This period of UK politics provides an interesting case study, in which MPs change their political stance en masse to match that of the public. Observing the differences between MPs who change their stance and those who do not, may provide insights into how changing political views manifest in language, and will require investigation using methods that look at language change. Another advantage of this data is that MPs are public figures about whom we can gather extensive metadata, unlike anonymous users in web communities. We also have some idea of their political views, at least on a public facing level. There are also key events that happened between 2015 and 2019, primarily the referendum itself, which can be used as points around which to observe the language of MPs. Seeing whether these events act as change points in language usage could be very informative.

4.2.2 Building the Corpus

To build the corpus, we downloaded the XML versions of the Hansard records from the UK Parliament API [6]. All House of Commons debates were gathered along with metadata. We put all this information into a database with the following tables.

• **Contributions** Every utterance in the dataset. For each contribution we gathered:

1. The text of the contribution.
2. The debate this contribution was from.
3. Whether or not it was a question.
4. If it is an answer, to which question does it refer?
5. Topic of the question.
6. Section in which the contribution was made. Usually this is a topic.

[6] `https://data.parliament.uk`

7. Contribution Type – Code allocated by api, e.g. 'Start Answer'.
8. The tag for the section, e.g. 'Debated Motion'.
9. If it was a question, the department to which the question was addressed.

• **Debates** Contains each debate, including its date and Hansard File.

• **Member Constituency** Keeps a record of constituencies for each member, with an entry for each election.

• **Member Party** Similar to above, but records the party memberships of each member.

• **Member Stances** Records a set of votes and political positions for each member. Votes were gathered using the UK Parliament Vote API [7].

1. Referendum Stance of MP [8].
2. The way their constituency voted in the 2016 EU Referendum [9].
3. Vote in Theresa May's deal (First deal).
4. Vote on the Benn Act, forcing the Government to prevent a no-deal Brexit.
5. Vote on Boris Johnson's deal, that went through.

• **Members** A table of all Members of Parliament with some metadata.

1. ID numbers. Several exist and are used in different parts of the API.
2. MP names.
3. Current Constituency, as of the point in time gathered.
4. Current Party, as of the point in time gathered.
5. Start and end date of being an MP.

[7] `https://votes.parliament.uk`

[8] Based on a list published by Politics Home: `https://politicshome.com/news/europe/eu-policy-agenda/brexit/news/dods-people/76451/interactive-map-every-mps-eu-stance`.

[9] The vote counts were not published per constituency, however people have predicted the numbers based on statistics from voting areas, as well as numbers from constituencies for which results were published. The data, and a more detailed explanation of how it was gathered, can be found here: `https://commonslibrary.parliament.uk/parliament-and-elections/elections-elections/brexit-votes-by-constituency/`.

4.2.3 Limitations and Caveats

This data is not without its limitations. The API is very limited and the formatting of the data is inconsistent and messy. Artefacts are therefore likely to exist, and though all known ones were removed, it is possible that there are more we did not catch. Text entries originally contained tagged meta information, such as actions. However, the tagging of these actions is not consistent enough to be completely sure they have all been removed. Another issue is that sometimes Hansard records contain transcription mistakes, or inaccurate accounts of what was said, which can cause problems when using it as a corpus [Mollin, 2007]. Updated Hansard editions are frequently published with errors corrected, and where possible we ensured that we always kept the most recent. This does, however, mean that the latter entries in the corpus may contain uncorrected errors.

The integrity of the corpus was tested by sampling four debates from each year and manually checking the correctness. This helped to rule out the probability of a sudden format change. We also checked the MP data manually, making sure that names were recorded correctly in cases where an MP might have more than one name (e.g. “Theresa May” and “Prime Minister”), or where a single name might have more than one MP (e.g. “David Cameron” and “Theresa May” both being “Prime Minister”). After these checks, we were confident that the data was correct to a level where it could be useful. We must be aware when analysing the data, however, that there could still be a level of noise, e.g. occasional missed meta-tags.

This dataset also provides some challenges for looking at community language change. One possible issue is that, compared to other communities, there are fewer members and relatively static membership. Another problem is that the amount of text per user is very asymmetric: ministers and those in government are far more likely to speak than backbenchers. There is also a challenge of data not being totally continuous. Parliament takes regular recesses, which creates gaps in the data. For example, in our entire corpus, there is never an entry in August. This could cause sliding window approaches based on time to behave unexpectedly.

4.2.4 Creating Groups

In this work we want to find out if different ideological groups or sub-communities use language differently. Within parliament, the most obvious grouping to look at is political party. This is a useful grouping because members can be definitively labelled and the groups have parallels to ideological belief, though of course there will still be a broad range of ideology within each group.

Other groupings are based on stances of MPs in the EU Referendum. 'Leave' and 'Remain' could be useful sub-communities to look at in contributions relating to the EU. These grouping could also be made more specific. For example, we can look at four groups based on MP Referendum stance and the vote of their constituency. This allows us, for example, to see if remain MPs whose constituencies voted leave change their language after the referendum, differently from remain MPs for remain constituencies.

One can also look for smaller, more specific groups, for example members of the anti-Europe "European Research Group" or the pro-EU "Rebel Alliance" of Conservative MPs. These may provide more specific groups containing fewer members, and will challenge our analysis methods' ability to work with much more limited data.

All of these groupings are just examples. We wish for our methods to be applicable to any group, and not limited to parliamentary data.

4.2.5 Looking at Specific Topics

We will sometimes want to look at language on a specific topic. For example, the 'Leave' and 'Remain' groupings make most sense when related to discussion involving the EU. To separate contributions on a specific topic, we use 'filter functions' to narrow down the contributions to look at. The most simple filter function is keeping only contributions that mention a specific term or a set of keywords. Analysis using this technique is not too different from looking at collocations and concordances in linguistics.

4.2.6 Preprocessing and Tokenisation

For the analysis, we of course had to preprocess and tokenise all of the text. In preprocessing we removed excess spacing, newlines, and made everything lowercase.

Making all text lowercase was done to remove the possibility of any word's counts being split between a capitalised version (e.g. at the beginning of a sentence) and the normal, lowercase version. This may cause problems with losing the ability to distinguish proper nouns, especially when a name is also a common word. The most notable example of this would be "May", the surname of UK Prime Minister Theresa May. Names are not especially common in this corpus however, as names of MPs are only used by the Speaker of the house, by convention. The spaCy tokeniser [Honnibal and Montani, 2017b] was used for tokenisation.

4.2.7 Sliding Windows

Because we want to look at language change over time, we split the data up into time windows – sets of contributions between consecutive date ranges. We used two types of window in this work: time and contribution windows. Time windows consist of all contributions between two specific dates. These windows can be moved along at regular time intervals, for example every month. Contribution windows on the other hand, contain a fixed number of contributions. The window is slid along by a specified number of contributions.

Both forms of window have their benefits. Time windows are more intuitive, and make it easier to compare between parallel windows from different groups, because time passes the same for everybody. Contribution windows, however, are useful because they are always the same size, and do not suffer from inconsistent window sizes when faced with gaps in the data.

For much of this work we used overlapping sliding windows. This means that we moved windows along at time intervals smaller than the window size. For example, we might collect a year of contributions in each window, and slide the window forward a month at a time. Overlapping windows have two main advantages. Firstly, they mean that we can use more data for each window, while still looking at many time points. Secondly, they mean we do not need to make assumptions about our corpus by splitting our data into epochs manually. For example, there is no reason why we should look at data from each discreet month, or year, separately, compared to any other arbitrary means of splitting the data.

Table 4.1: Basic Meta Features of the corpus, including statistics for the subset that contains only EU mentioning contributions.

	All	EU Mentions
Number of Contributions	275,066	58,221
Number of Words	46,874,350	12,580,593
Number of MPs	749	743
Users w/ >50 Contributions	709	291
Median Words/Contribution	84	95
Median Words/MP	47,074	9,362
Median Contributions/MP	253	36

4.2.8 Meta Analysis

We performed a meta analysis to understand how data in the corpus was distributed. First we gathered some basic statistics for the entire corpus as well as for a subset of the corpus that mentions the EU. To gather this subset we kept all contributions that contain a number of key words and phrases[10], as well as contributions with section headings (essentially debate topics) that contain certain terms. Table 4.1 outlines the basic metadata for the dataset.

As evident in the table, this corpus contains a large amount of text, given a fairly limited time span, but has a relatively low number of unique speakers. Most MPs have more than 50 contributions, meaning that there is a reasonable amount of text for each speaker. Another thing to note is that the number of MPs is higher than the 650 MPs sitting at any given time because of new MPs elected over the course of our timespan.

When looking at just EU text, there is a disproportionally large number of words for a relatively small number of contributions. EU Contributions also tend to be longer. Most speakers in the corpus are also present in the EU subset, meaning that they mention the EU in at least one contribution. This could suggest that the subset is fairly usable despite it being substantially smaller.

Figures 4.2 to 4.4 show the distribution of several meta variables. They all show that many of the words and contributions in this corpus belong to a small number of speakers, and that the majority of contributions are very short, with a few extremely long ones. Dealing with such an imbalanced corpus is a challenge, but an inevitable problem of a dataset where certain speakers, e.g. government ministers, will always contribute more than others. We will need to be aware that we may need to discard a

[10]The full list of which can be found in the supporting materials.

Table 4.2: Basic Meta Features of the corpus, for four groups: Conservative, Labour, Remain, and Leave.

	Conservative	Labour	Remain	Leave
Number of Contribs	169,532	70,168	189,880	64,046
Number of Words	25,962,883	14,069,725	33,195,314	9,882,757
Number of MPs	363	288	477	156
Users w/ >50 Contribs	352	263	456	148
Median Words/Contrib	82	88	86	80
Median Words/MP	51,994	40,612	57,221	44,671
Median Contribs/MP	312	203	293	275

large amount of data to accommodate for this.

If we look at contributions below the 75th percentile of contribution length (Figure 4.5), the lengths are fairly normally distributed, except for a mini peak which is presumably reserved for very brief contributions. A similar story is true for contributions per user (Figure 4.6), except without a second peak.

EU Contributions seem to be distributed similarly to the general corpus. Figures 4.7 to 4.9 show the distribution of contribution length, words per MP, and contributions per MP. The main difference from the overall corpus is that the number of words and contributions per MP tends to be lower. This is more evident in the plots of the distribution above and below the 75th percentile, as shown in Figures 4.10 and 4.11.

Figure 4.12 plots the cumulative number of contributions over time for the corpus. From this we can see that there is a steady flow of contributions across the time span. There are parts where the line flattens out temporarily. This is due to parliamentary recesses, and events such as elections, where parliament is suspended.

4.2.9 Meta Analysis of Groups

Table 4.2 shows the basic meta-statistics for four groups: Conservatives, Labour, Leave, and Remain. Labour and Leave are much smaller groups than Conservative or Remain. However, the data is similarly distributed for all groups. The median words per MP is lower for the two smaller groups, probably because the individuals who speak the most are Government Ministers, all of whom are Conservative (in this date range), and the majority of whom supported remain. Despite this, the smaller groups still have a substantial amount of data. The distributions also follow the same trends as the general corpus.

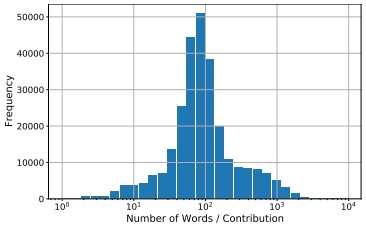


Figure 4.2: Histogram of Words per Contribution for all contributions.

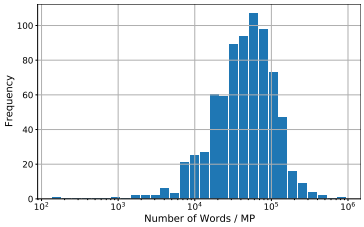


Figure 4.3: Histogram of Words per MP for all contributions.

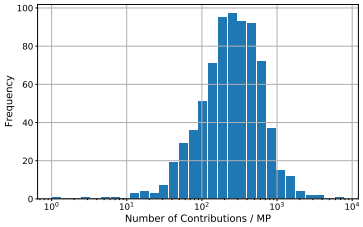


Figure 4.4: Histogram of Contributions per MP for all contributions.

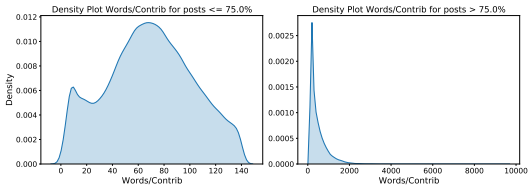


Figure 4.5: Distribution of Words per Contribution, above and below 75th percentile, for all MPs.

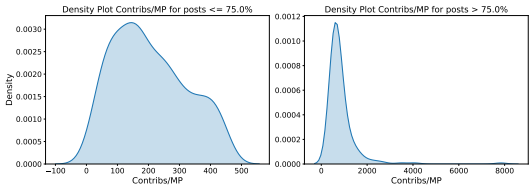


Figure 4.6: Distribution of Contributions per MP, above and below 75th percentile, for all MPs.

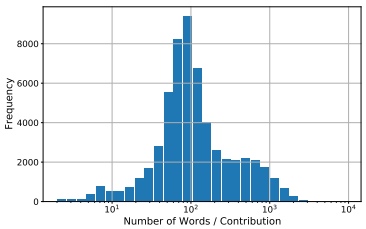


Figure 4.7: Histogram of Words per Contribution, for EU-mentioning contributions.

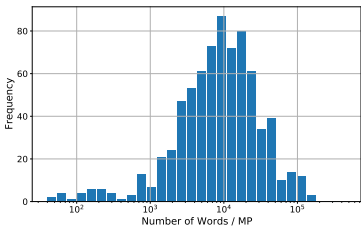


Figure 4.8: Histogram of Words per MP, for EU-mentioning contributions.

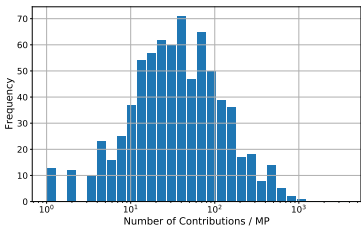


Figure 4.9: Histogram of Contributions per MP, for EU-mentioning contributions.

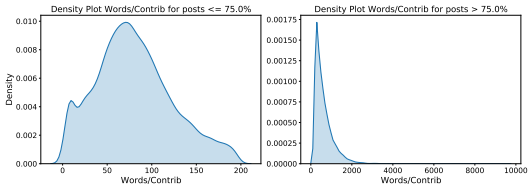


Figure 4.10: Distribution of Words per Contribution, above and below 75th percentile, for EU-mentioning contributions.

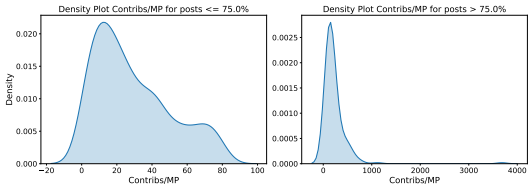


Figure 4.11: Distribution of Contributions per MP, above and below 75th percentile, for EU-mentioning contributions.

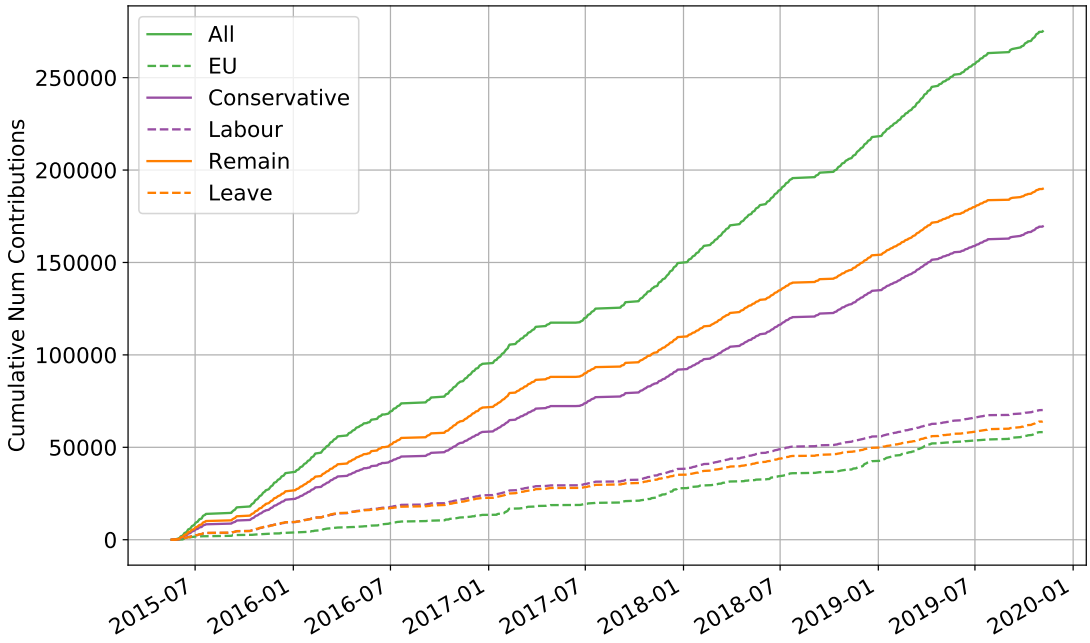


Figure 4.12: A plot of the cumulative number of contributions for All contributions and EU Mentions, as well as for a selection of groups.

4.2.10 Looking at the Top Speakers

To conclude our meta-analysis, we looked at the MPs with the most contributions. The ten most highly contributing MPs are listed in Table 4.3. This highlights some of the problems discussed earlier.

Eight of these MPs are Conservative, compared to one Labour MP, who in this case is the Deputy Speaker, who does not voice political opinions and presides over certain proceedings. Similarly, there are six remain MPs, compared to two leave. Even within the top 50 MPs, there are only four Labour members and thirteen leave. Slightly more balanced is the voting of each MPs constituency. Six of the top ten represent remain voting constituencies compared to four who represented leave. When looking at the top 50, 20 held remain seats and 30 held leave seats.

Table 4.3: A table of the 10 MPs with the most contributions, showing the number of contributions for each, alongside their party, referendum stance, and stance of their constituency.

Name	# Contribs	Party	Referendum Stance	Constituency Leave %
John Bercow	11,817	Speaker	Unknown	48.9 (Remain)
Theresa May	7,837	Conservative	Remain	45.0 (Remain)
Andrea Leadsom	4,076	Conservative	Leave	53.3 (Leave)
Chris Grayling	3,480	Conservative	Leave	47.8 (Remain)
Sajid Javid	2,714	Conservative	Remain	55.4 (Leave)
David Lidington	2,515	Conservative	Remain	51.8 (Leave)
David Cameron	2,269	Conservative	Remain	46.3 (Remain)
Jeremy Hunt	2,161	Conservative	Remain	40.7 (Remain)
Greg Clark	1,989	Conservative	Remain	44.6 (Remain)
Lindsay Hoyle	1,768	Labour	Unknown	56.7 (Leave)

These top ten MPs make a considerable number of contributions. The top ten made $\sim$41,000 contributions and the top 50 made $\sim$89,000. This goes to show how a relatively small number of MPs do a large proportion of the speaking. Another point worth noting is that the MP with most contributions is the Speaker, who should not be included in any of the groups, as their job is to remain neutral and oversee the proceedings of the house.

These imbalances will have to be carefully considered when dealing with this type of data. Developing methods that are robust to this problem will be useful when analysing many communities. Even online where almost everybody has an equal ability to speak, there still end up being certain members who dominate conversation.

4.3 Keywords and Wordclouds

The first method we applied to our data was a keyword analysis. This involved finding keywords in the corpus for different groups and then comparing them by plotting them inside word clouds. We could then look at the collocates of these words to assess what they can tell us about the groups based on their context in sentences. This method is less for looking at language change over time, and more for identifying the words that distinguish groups from one another. If we can identify words that are characteristic of certain groups, we can use them as features when looking at diachronic change.

4.3.1 Background

Looking at key words in a corpus is a common technique in corpus linguistics and has many potential applications, from visualising corpora, to finding topics [Kilgarriff, 2001]. There are various measures for finding keywords. The most basic might be to find the most common words in different parts of the corpus, ignoring common stopwords (e.g. 'the', 'a'). The problem with this approach is that it will consider words that are inherently common as being key, which is not particularly helpful. For example, if the word "speaker" is common amongst all groups, it is not a useful keyword.

We would be much more interested in a method that considers words key if they appear significantly more in one subcorpus than another. More complex methods for finding keywords, may measure statistical significance or effect size of a word between corpora. Methods for evaluating significance, e.g. Chi-squared and Log-likelihood, will tell us which words have the most statistically significant difference between corpora [Rayson and Garside, 2000]. Effect size measures, on the other hand, will tell us which words are the most different between two corpora. Log-ratio is an effect size statistic which measures how many times more frequent a word is in one corpus than the another by calculating the binary log of the odds ratio of the word's relative frequency in each corpus. These can be used in conjunction. You may, for example, find the n most significantly different words using a significance measure, and then sort them by effect size. For more background on keyword analysis, refer back to Section 2.2.1.

Word clouds are a commonly used method for visualising corpora. They are not often used in a scientific context, but as a tool for displaying common words in a text, they can be useful. They can also be used to show other features such as semantic domains and parts-of-speech.

4.3.2 Keywords

First we calculated the keywords for each group, across the entire corpus. As we were interested in debates surrounding the European Union and Brexit, we took the subset of contributions for each group that mentioned a set of Brexit related terms and compared the keywords of this subset against the set of all contributions that do not mention these terms, irrespective of group. Therefore, a keyword is a word that was used far more by this group when talking about the EU than in general parliamentary contributions that

are not about the EU.

We considered keywords as being words with a Log-Ratio of more than 1. This threshold meant that we only kept words that appeared at least twice as much in the current group as in the reference corpus. It may have been preferable to use a significance threshold, but log-likelihood values are not comparable when comparing to corpora of different sizes, unlike log-ratio which means the same thing for any corpus. To remove infrequent words that only appear very occasionally, and therefore appear as deceptively key, we only considered words keywords if they had a frequency of more than 10 across the entire corpus. This cut-off was chosen as it was high enough to exclude most words that only appeared in a single contribution, but not so high that words that were spoken a lot in a single debate would be discarded. It also had the added benefit of removing spelling mistakes, which would always come up as key without this measure in place.

As well as finding all the keywords for each group across the entire corpus, we also found keywords over time. We did this by iterating through each time window (Section 4.2.7) and finding the keywords in exactly the same way as described above, except only looking at contributions from within that window. While we will not visualise these keywords in this section, they can be used as features in future sections. This will help us determine if the change in these keywords can be used to observe how language changes in these groups over time.

4.3.3 Visualising Keywords for Different Groups

In order to visualise the keywords of groups in our corpus we used Word Cloud Venn diagrams. These were created using a python package[11] which produces Venn diagrams where each circle is a wordcloud. Figure 4.13 shows an example of a word cloud Venn diagram from the Hansard data. These were keywords gathered from all posts, across the entire time span.

Each circle of the diagram holds the keywords for a specific group. Where a word is a keyword of multiple groups, it appears in an intersection of the circles to which it belongs. For example, the word ‘Brexit’ appears in all three groups, therefore it is in the central, 3-way intersection. Whereas, the word ‘vassal’ appears only in the group of

[11] `https://pypi.org/project/matplotlib-venn-wordcloud/`

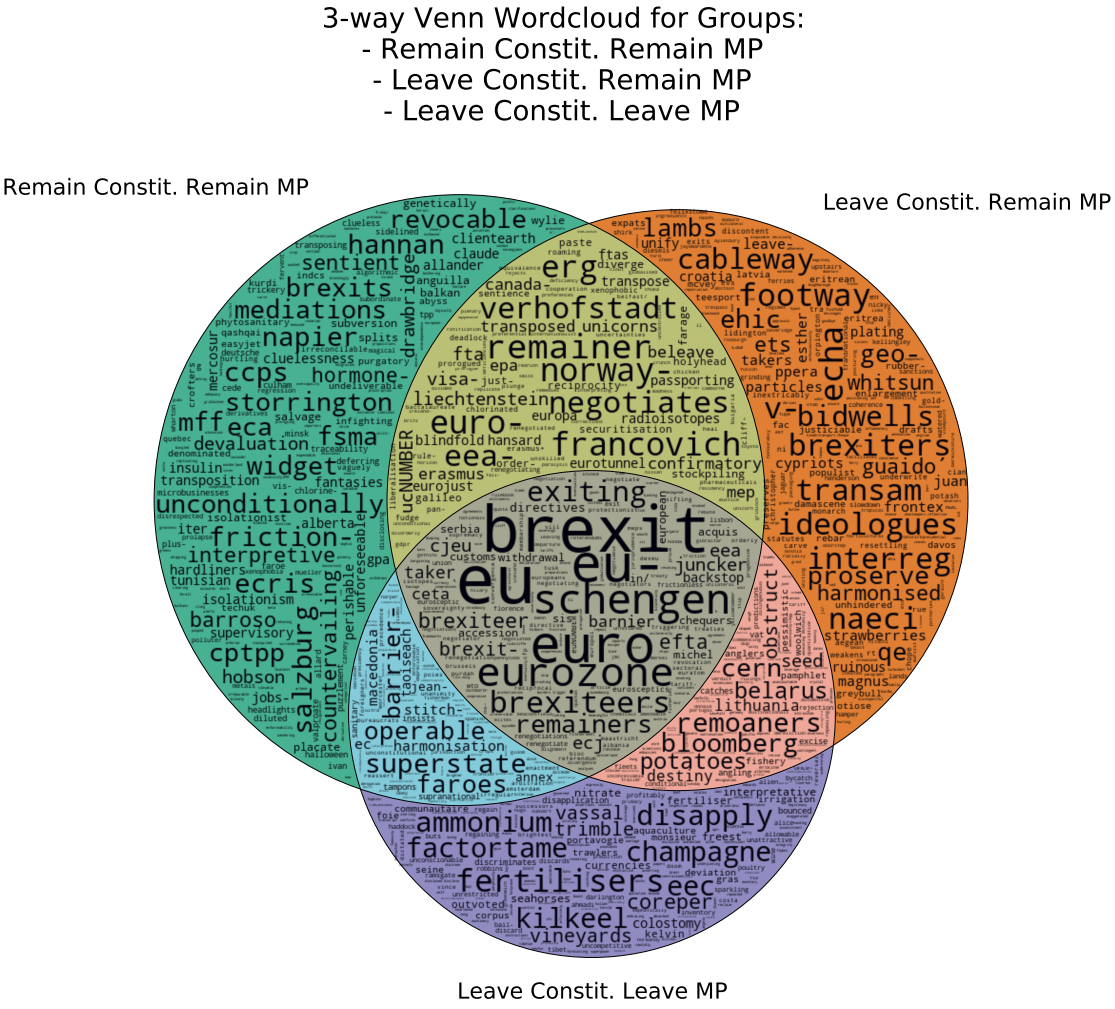


Figure 4.13: Wordcloud Venn diagram for 3 groups in our corpus. These groups are based on MPs support of leave/remain in the 2016 EU Referendum and the way their constituency voted. While this example shows these specific three groups, it would be possible to create a word cloud Venn for any combination of groups, although it becomes difficult to parse for more than three groups.

leave-supporting MPs who's constituencies voted leave.

From this plot, we can learn about the way different groups use language. For example, there are certain words all three groups use, such as "Brexit", "EU", and "withdrawal", which may be considered neutral Brexit-related terms. Other words are more group-specific. Leaver MPs from Leave constituencies, for example, overuse words such as "vassal" and "superstate", playing up fears that Britain would be consumed by the EU and lose its independence and identity, should it stay in the European Union. Remainer MPs from Remain constituencies, meanwhile, overuse the term "drawbridge", referring to Britain pulling up its metaphorical drawbridge and isolating itself from the world. Remainer MPs from both Leave and Remain constituencies shared certain keywords, such as "visa-free" and "Erasmus". These keywords being shared between these groups possibly shows examples of discussion where MPs are aiming for assurances on certain details of a post-Brexit deal, rather than resisting Brexit outright. On the other end, both Leave and Remain MPs from Leave Constituencies overuse the terms "remoaner" and "obstruct", both words used to complain about perceived attempts of Remainers to overrule the referendum result.

This analysis does not say anything about diachronic language change. However, we believe it can be useful in guiding analysis by highlighting words that distinguish certain groups. Also, while we do not show it here, one could quite easily produce these Venn wordclouds at any time step. In this case, the method could serve as a useful visualisation tool for performing a qualitative analysis of language change.

4.3.4 Concordances

Concordance is a widely used technique in corpus linguistics for showing the context in which words appear. Here we will provide a demonstration of how we can use the insights from this analysis to learn interesting things about the different groups. A word that stuck out in the wordcloud above is 'drawbridge'. We wanted to know more about the context in which it was used in parliamentary discourse. To investigate this, we created the concordances of the word shown in Figure 4.14.

The concordances show that MPs on both sides of the debate use the phrase, although Leave-supporting statements, e.g. "we are not pulling up the drawbridge", tend to be on the defensive. This phrase is clearly a common argument used by Remainer

```
Displaying 25 of 53 matches:
rs that shows that we have not pulled up the drawbridge . let me again quote fabien miclet from live
m result as a vote for the uk to pull up the drawbridge . we will remain an open , tolerant country
does not mean we are advocating pulling up a drawbridge . in certain areas , whether allowing the re
ver , but once in power , they pulled up the drawbridge and were nervous about the challenge they fa
hat does not mean that we are pulling up the drawbridge and ending all immigration . in fact , it is
er will know that , as he plans to raise the drawbridge into england through raising fees , in wales
 now , but this will not mean pulling up the drawbridge . we will operate the immigration system in
f open - door immigration and pulling up the drawbridge , it seems to me that there is huge scope fo
 facing . now is not the time to pull up the drawbridge . now is more than ever the time to open the
, build a wall around the uk and take up the drawbridge . it fundamentally fails to take account of
s it makes clear , we are not pulling up the drawbridge or building some imaginary sea wall down the
 can not just close our eyes and pull up the drawbridge ? before my right hon . friend gets on to de
ment without saying that we will pull up the drawbridge . we need to have a balanced approach to imm
 immigration to this country and pull up the drawbridge . rather , i take it as a signal that the br
e referendum result as a vote to pull up the drawbridge . the united kingdom will remain an open and
from free trade should not be pulling up the drawbridge behind them and denying those benefits to de
nomy . the thought that we would take up the drawbridge and prevent people from coming to participat
 of the world , or do we want to pull up the drawbridge ? some people want to pull up the drawbridge
drawbridge ? some people want to pull up the drawbridge and to let the world get on and pass us by .
nto a closed - minded attempt to pull up the drawbridge . this country is at its best when it is ope
ther parts of the union , not to pull up the drawbridge and say , " we 've got what we want . now we
is the best approach ? should we pull up the drawbridge or co - operate with our neighbours ? the la
 part of a xenophobic society that pulls the drawbridge up behind us . our universities have express
. stumbling out of europe and pulling up the drawbridge would serve only to harm our position and in
ng we could do is to walk away , pull up the drawbridge and say it is all too difficult . though an
```

Figure 4.14: An example of the concordances of the word ‘drawbridge’ in our corpus.

MPs, to comment on Britain’s perceived isolationism. It is interesting that a medieval metaphor has been chosen to demonstrate this: perhaps its archaism implies Britain is being backwards or old fashioned. It would be interesting to further investigate the usage of medieval language in Brexit discourse – we have already seen the word “vassal” being a keyword for Leaver MPs. This is an example of the kind of analysis that could be aided by this technique.

4.3.5 Keywords as Features

For much of the analysis in the future sections, we used Keywords as features. To create a keyword feature set, we looped through the data, window by window, as described in Section 4.2.7. For each window, we found the keywords as defined in Section 4.3. Once we had a set of keywords over time, we created a keyword (KW) feature vector for each contribution. This vector contained a frequency count for each word that appears as a keyword in any window. These vectors can be summed to create a vector for each window. Before use, all counts are normalised by text length.

4.3.6 Concluding Remarks

In this section we have demonstrated how keyword analysis can be employed to compare the language of groups in a community. While we did not cover diachronic

analysis, we did highlight some words that certain groups within parliament overuse compared to others. These words could next be tracked using diachronic methods. Keyword analysis is a simple technique, that can yield quick and easy to understand results. This makes it perfect as an initial form of analysis to precede more complex diachronic language change techniques.

4.4 Diachronic Word Embeddings

In NLP, there has always been a need to find ways of representing words. Many popular methods rely on the idea of distributional semantics [Firth, 1957]; the idea that similar words appear in similar contexts. More recently, neural methods have become popular such as Word2Vec [Mikolov et al., 2013]. These methods create distributional word embeddings for each word. An embedding is a dense vector, and semantically similar words will have similar vectors based on their context in the training corpus. Along with Word2Vec, many other methods have been developed to perform similar tasks [Pennington et al., 2014, Devlin et al., 2019a].

Previous research has used Distributional Word Embeddings to look at how language changes over time [Kutuzov et al., 2018]. These techniques can be used for tasks such as investigating cultural or semantic shift [Hamilton et al., 2016a, Kulkarni et al., 2015], and they provide different insight from more simple frequentist approaches, such as plotting word frequency over time. These methods involve training embedding models at separate time steps, and then comparing a word's vector from one time step to another, or looking at a words neighbours over time. For a more detailed overview of word embeddings, and semantic shift, refer back to Sections 2.1.2 and 2.3.1.

In this chapter, we will observe language change by looking at the nearest neighbours to a word over multiple time steps. We will also identify the words that changed the most using the method described by Gonen et al. [2020] for measuring usage change. Specifically, the method we employed was as follows:

1. Split the corpus into windows of 365 days, producing four windows.

2. Train a Word2Vec model [Mikolov et al., 2013] on each window.

3. For each word:

(a) Calculate the 1,000 nearest neighbours for each time window. Neighbours were found using cosine similarity.

(b) Get the intersection of the list of nearest neighbours, for each subsequent window.

By doing this, we can see which words changed in usage the most at each time window, compared to the last. Words which have smaller intersections of neighbours with the previous window will have changed more in usage than those with a larger intersections.

Various challenges present themselves when using this method for our dataset. Firstly, in the past semantic change has mostly been used to look at long term shifts, over decades or centuries, although there are exceptions [Del Tredici et al., 2019]. The shorter time range means that shifts will mostly be topical, and it is possible that the time range is too short to capture many changes. The second challenge is the quantity of data. Our corpus is not enormous, largely due to its limited time range. This means that models trained on our data may not be especially good, particularly for rare words. The problem is exacerbated when the corpus is further split into time windows. There are ways of mitigating this problem – for example, we could ignore rare words when calculating neighbours.

4.4.1 Training Diachronic Word Embeddings on Hansard

Using diachronic word embeddings, we hope to get an impression of how some of the keywords surrounding Brexit change over the timespan of our corpus. The first step in achieving this was training a language model at multiple time intervals. Initially, we used non-overlapping 365 day windows. One of the main considerations in this process, is that a small window means the model will not be trained on as much data, and as such may not be as good a model. Using a year as the window size ensures a balance between data quantity, and number of windows. We used non-overlapping models because it is how similar models have been trained in the past.

Because we only intend on comparing the neighbours of words in each time window, rather than the vectors themselves, we do not need to align the models in any way. This significantly reduces the computational complexity of the process.

To sanity check our embedding models, we plotted the neighbours of several brexit-related keywords over time. Some of these words have logical changes that we expect to see. For example, words like 'single' and 'common' would not be expected to have the same level of EU link in 2015 as they do after the 2016 referendum, when they were used as part of 'single market' and 'common market'. So the first way we evaluated our embeddings was by looking at the nearest neighbours (closest word vectors) for each of these keywords at each time step. This would help us see if the diachronic embeddings are capturing semantic change as we would hope.

Once we have verified the soundness of our embeddings, we can learn about the semantic shifts of words within parliamentary debates. This will be achieved by identifying the words that change the most at each time step, using the process described above. By looking at the neighbours of these words, and when the change occurred, we can make inferences about the reasons behind the change and how it manifested. Because we are using year long windows, this probably will not help us identify particular events, but it may still give an impression as to the years in which language shifted. Looking at smaller, or overlapping, windows may help to identify more granular stages of change in the future.

This analysis will initially be run on the entire corpus, to observe semantic shifts in parliament as a whole. Following that, the investigation will look at individual groups, to see whether different words change dramatically for different groups. If this is the case, we will be able to look at these words and gain insight into the groups. A potential pitfall of looking at groups is that it will make the dataset much smaller. There are potential ways around this. One would be to use overlapping windows to increase the number of words per window. Another would be to pre-train the embeddings on all contributions before training on a specific subset of MPs.

4.4.2 Static Embeddings

To begin with, we trained static word embeddings on our corpus. For this we used Gensim [Řehůřek and Sojka, 2010] to train a Word2Vec model on all our text with default parameters, and a vector size of 300. We only looked at the embeddings of the most frequent 10,000 words. This was due to the corpus being fairly small, and uncommon words being given poor embeddings due to a lack of data. We then looked at

Table 4.4: Table of nearest neighbours for four selected words relating to Brexit in the static word embeddings trained on our corpus.

Word	**Neighbours**				
referendum	election	vote	elections	brexit	leave
	voted	votes	article	parliament	negotiations
brexit	referendum	exit	deal	vote	eu
	austerity	negotiations	devolution	backstop	outcome
immigration	migration	asylum	welfare	fisheries	detention
	border	trade	visa	migrants	sanctions
leave	stay	remain	referendum	leaving	left
	exit	lose	leaves	go	vote

the nearest neighbours of a selection of words relating to the EU and Brexit. Neighbours were found by looking for the vectors with the highest cosine similarity to the query word's vector.

Table 4.4 shows the ten nearest neighbours of four example words that relate to Brexit. More than anything, we performed this step to sanity check the model, by making sure that the embeddings it produced made sense.

On inspection, they do seem to make sense. 'Referendum' is close to words relating to elections and votes, as well as 'brexit', the widely adopted name for the EU referendum. Similarly, 'brexit' neighbours words that are significant to the referendum, e.g. 'eu', 'outcome', and 'backstop'.

An interesting thing to note about the neighbours for 'immigration' is that 'fisheries' lists among them. Fisheries were widely discussed around Brexit, but not so much beforehand. It would be interesting to see how this word changes in our diachronic analysis, to see when this neighbour is prevalent.

Overall, our embeddings seem sensible. It is worth noting that these neighbours are drawn only from the most common 10,000 words in the corpus. We think that this suits our purposes fine, as we are mainly interested in common, subject specific vocabulary. We could have remedied this problem by starting with a pre-trained word-embedding. We will look into employing this technique in future work.

4.4.3 Diachronic Embeddings

Next we trained embeddings each year beginning in May. This involved a similar process as before, training a Word2Vec model for the contributions of each year

Table 4.5: Table of the nearest neighbours for each year long window of word embeddings trained on our corpus. New words highlighted in bold.

	brexit				
May 15 - 16	eu	vote	leave	european	election
May 16 - 17	**referendum**	eu	**negotiations**	leave	**trade**
May 17 - 18	**exit**	eu	trade	referendum	**union**
May 18 - 19	**deal**	referendum	vote	**backstop**	**prime**
	immigration				
May 15 - 16	welfare	criminal	justice	migration	sanctions
May 16 - 17	**foreign**	**eu**	**prime**	**brexit**	**movement**
May 17 - 18	trade	eu	**tax**	**legal**	**customs**
May 18 - 19	trade	**justice**	**fisheries**	**migration**	tax
	single				
May 15 - 16	every	one	one-	two-	union
May 16 - 17	union	**eu**	**european**	**labour**	**market**
May 17 - 18	union	**customs**	labour	**common**	eu
May 18 - 19	common	customs	**every**	union	eu

individually. This of course meant that the embeddings were trained on less data, which will mean lower quality embeddings[12]. We then looked at how the neighbours of certain EU related words changed over time. For now, this is mainly to show how this method can be useful even over such a short time span. Table 4.5 shows the changing neighbours of three words that highlight the usefulness of this method.

The word 'brexit' is fairly stable in its neighbours over time, with neighbours all relating to the EU referendum. Although it seems stable, there does appear to be a slight shift in focus from neighbours relating to the vote in general (e.g. 'vote', 'election'), to words relating to the aftermath of the referendum (e.g. 'negotiations', 'trade'). There are certain words we might expect to be similar, such as 'hard', 'soft', and 'clean' that never feature in the top five. But it would be interesting to see how the similarities of these words to 'brexit' change over time. The main indication of change here, is the introduction of 'deal' as the nearest neighbour in the last window, when the Brexit deal was being debated in parliament. Along similar lines, 'backstop' becomes a near neighbour in this window.

'Immigration' changes subtly over time. In the first window, before the referendum, its neighbours do not particularly relate to Brexit, with words like 'migration'. 'Criminal' and 'welfare' are also neighbours, possibly because they are also followed

[12]Data quantity is a significant limitation of this technique, and in future work it would be helpful to thoroughly test the boundaries of what constitutes enough data.

commonly by the word 'system', but also could be topics that are discussed alongside immigration. In windows following the referendum, its neighbours become far more EU related. Words such as 'Brexit', 'EU', and 'movement' come up. This suggests, unsurprisingly, that, after the vote, much of the immigration conversation was focused around Brexit and the EU. This is further suggested by the presence of the word 'fisheries' in the final window – a word that has not really got anything to do with immigration, other than it also being a hot topic of conversation during Brexit debate.

The final word we show here is 'single'. This was chosen as it is a common English word that has a particular meaning in EU discourse when discussing the 'single market' – a major talking point of the debates around the UK's plans for Brexit. This is reflected in its change. In 2015-2016, 'single's neighbours are as one would expect in non-EU related conversation, for example, 'every', and 'one'. There are some neighbours that possibly link to the EU meaning, e.g. 'union'. Over the subsequent windows, however, the neighbours are increasingly EU related. Words like 'common', 'EU', and 'customs' become neighbours. This change is not massively significant, but given how short the time span is, it provides us with an idea of how this word's usage changed.

4.4.4 Finding the Most Changing Words

Having shown so far only words with a predictable change, the next step is to find words that change dramatically over time. This could guide linguistic analysis of parliamentary language change by highlighting words for linguists to investigate, possibly with more traditional, corpus driven techniques. We will identify such words using the method introduced by Gonen et al. [2020], and described in Section 4.4.

Table 4.6 shows the 20 words that changed the most in usage between each time window. Only words that appeared in the entire dataset at least 100 times were considered. This was to remove rare words, and focus on widely used words that changed in meaning[13]. Even with this limit, many of the highlighted words are difficult to interpret. This is not too much of a problem, however, if the technique is used alongside manual analysis.

It is immediately evident that some of the words, such as "dog" are not particularly

[13]The value of the threshold was chosen intuitively after testing several values, with 100 achieving a balance of excluding very rare words while also allowing relatively specific terminology.

Table 4.6: Table of the most changing words between each subsequent pair of windows.

May 2015-16 to May 2016-17 *(Window 1 - 2)*				
google	dog	customs	style	similarly
plain	bomb	e-	strikes	rbs
tv	managing	exit	grammar	independently
trading	supreme	s.	brexit	smith
May 2016-17 to May 2017-18 *(Window 2 - 3)*				
retained	osborne	selection	tower	super-
principal	no-	salisbury	radio	bbc
shipley	similarly	privatisation	wear	s.
leigh	chemical	philip	semitism	continually
May 2017-18 to May 2018-19 *(Window 3 - 4)*				
offensive	grieve	permanent	charter	moreover
bone	principal	zero	and-	white
basically	virtually	letwin	overnight	card
mixed	no-	presumably	furthermore	meanwhile

obvious. This may be a result of the size of the corpus creating low quality embeddings for relatively uncommon words. There will always be some noise in these results as the neighbours of a word may vary randomly. Further analysis is required to explain why words really change in usage. Table 4.7 demonstrates a choice selection of most changing words, showing each word's nearest five neighbours at each window.

Despite these odd words featuring, there are many that make perfect sense. For example, "Brexit" and "customs" change between the first two windows. This probably corresponds to Brexit discussion shifting from debate about the referendum, to dealing with the aftermath. This is suggested in Table 4.7 by the introduction of "trade" and "negotiations" into the near neighbours.

Many of the most changing words correspond to notable topics of discussion. For example, the word "strike" changes its usage. This probably links to a shift in usage from referring to military air strikes (a significant vote took place in 2015 on whether or not Britain should carry out air strikes in Syria), to more general usage. "Tower" and "Salisbury" are other words that change between the second and third windows. These link to two significant events: the Grenfell Tower fire[14], and the Salisbury Novichok poisonings[15].

Another word (more appropriately, token) that changed substantially was "no-".

[14] https://www.bbc.co.uk/news/uk-england-london-40269625

[15] https://www.theguardian.com/uk-news/2018/mar/07/russian-spy-police-appeal-for-witnesses-as-cobra-meeting-takes-place

Table 4.7: Table showing a selection of the most changing words, and their five nearest neighbours at each window.

	brexit				
May 15 - 16	eu	vote	leave	european	election
May 16 - 17	referendum	eu	negotiations	leave	trade
May 17 - 18	exit	eu	trade	referendum	union
May 18 - 19	deal	referendum	vote	backstop	prime
	customs				
May 15 - 16	gas	revenue	customs	trade	oil
May 16 - 17	customs	eu	trade	market	europe
May 17 - 18	european	trade	eu	border	euratom
May 18 - 19	european	backstop	deal	trade	agreement
	strike				
May 15 - 16	take	taken	vote	thing	industrial
May 16 - 17	strike	get	carry	be	hold
May 17 - 18	take	get	strike	give	be
May 18 - 19	negotiate	be	get	take	reach
	tower				
May 15 - 16	st	south	royal	city	(
May 16 - 17	(	st	hospital	royal	city
May 17 - 18	grenfell	blocks	fire	homes	cladding
May 18 - 19	tower	hospital	died	london	street
	salisbury				
May 15 - 16	st	south	royal	city	(
May 16 - 17	(	st	hospital	royal	city
May 17 - 18	grenfell	blocks	fire	homes	cladding
May 18 - 19	tower	hospital	died	london	street
	no-				
May 15 - 16	nuclear	air	where	anywhere	no-
May 16 - 17	nuclear	or	zone	create	fly
May 17 - 18	brexit	transitional	trade	no	great
May 18 - 19	no	without	bad	negotiated	great

Table 4.8: Table showing the neighbours over time for "sovereign", for the Remain and Leave groups.

	Remain				
May 15 - 16	devolved	member	secretary	democratic	european
May 16 - 17	democratic	united	leader	democracy	sovereign
May 17 - 18	european	nuclear	nation	democratic	democracy
May 18 - 19	nation	independent	european	democratic	british
	Leave				
May 15 - 16	european	rights	eu	our	nation
May 16 - 17	british	european	member	leader	kingdom
May 17 - 18	(	european	customs	international	rights
May 18 - 19	customs	eu	united	law	independent

This token changed from a usage relating to words such as "fly" and "zone" (possibly referring to 'no-fly zones') towards a Brexit-related meaning. This is probably due to it being a constituent part of no-deal - as in no-deal Brexit.

4.4.5 Comparing Groups

The intention of this chapter was to discuss various methods for looking at language change, and look at their suitability for comparing sub-groups within communities. So far, we have only used diachronic embeddings to observe changes in general language across all of parliament, over our time range. The approaches we have employed could easily be used to compare groups, however. By looking at a word's nearest neighbours over time for multiple groups, we may be able to learn more about how the language usage of certain groups contrasts to others. Similarly, the most changing words for two different groups may vary in interesting ways. This does present one major problem, however. Data will be stretched even thinner than before, potentially leading to weak embeddings, and too much noise to be useful.

To demonstrate how the method can be used to learn more about the language change of groups, we now present an example. Table 4.8 shows a comparison of the neighbours over time of the word "sovereign" for two different groups, Remain and Leave. There appears to be some difference in the nearest neighbours over time, with the Leave group having more clearly EU related neighbours from early on. That "sovereign" neighbours words such as "EU", "British", and "Kingdom" in the leave group, could suggest that the Leave group was more focused on the idea of nationalist sovereignty from the EU earlier

on. This could imply that the Leave group led on this particular discussion, as by the final window the Remain group has more EU related neighbours such as “independent”, “British”, and “nation”. The differences are not especially dramatic, particularly using only the top five neighbours, but this gives some idea of how this technique could be used in guiding analysis of groups within a corpus.

We also looked at the most changing words for different groups. However, we found the results to contain a lot of noise, with no clear explanations behind the differences between groups. This is most likely a problem caused by limited data, so while these methods are promising for group language change analysis, the data size issues need to be addressed before they are completely effective.

4.4.6 Concluding Remarks

This section has discussed the suitability of diachronic word embeddings for studying language change in groups. While we did successfully perform analysis that highlighted interesting aspects of language change, there were some key limitations that hinder its effectiveness. Most notably, a large quantity of data is necessary to make the method effective, though we did discuss some possible solutions for mitigating the problem. The results are also relatively oblique. While they can point you in the right direction, it is never completely clear why words are highlighted as highly changing, or what different neighbours mean. There is also lots of noise, with many words being highlighted for no discernable reason. One question worth asking, is whether the method contributes anything beyond more simple corpus linguistic techniques, such as looking at keywords, collocates and concordances. These alternative methods do not rely on large quantities of data and are more clearly understandable, although they are also less automated. Even so, diachronic embeddings serve as a more automated way of highlighting words that change in their meaning, and, given sufficient data, can be used to compare groups.

4.5 Variability-based Neighbour Clustering (VNC)

The next method we will detail is Variability-based Neighbour Clustering (VNC). This method is for identifying stages in diachronic corpora. We can use this to identify key stages in parliamentary debates. By comparing the stages of different groups, we can

learn about how the language of these groups varies over time. Unlike other methods, it will not tell us how a group's language changes so much as it will help us identify stages in which MPs used similar language. We can combine this with knowledge of real world events and gain a better understanding of which events possibly caused a change in language.

4.5.1 Background

Gries and Hilpert [2008] proposed a method called 'Variability-based Neighbour Clustering' (VNC) for identifying stages in diachronic corpora. This method was an adaptation of the Agglomerative Hierarchical Clustering algorithm, altered so that only elements that are chronologically adjacent can be clustered together. They used this technique to find stages in the development of specific language features.

The proposed technique was not a specific algorithm, but rather a general method which comprises of the following steps.

1. Begin with a cluster for each item being clustered. In our case, an item would be a feature vector representing a particular month, or time window.

2. Calculate the similarity between each cluster and the cluster that temporally follows (e.g. the next window). This can be done with any appropriate similarity metric, e.g. cosine distance (see Section 2.1.4).

3. The two most similar clusters are merged into a larger cluster.

4. Steps 2 and 3 are repeated until only one cluster remains.

5. As with hierarchical clustering, one can draw a dendrogram (see Figure 4.15) or pick a cut-off similarity to find the cluster (stage) of each item.

The applications that Gries and Hilpert [2008] discuss in their paper are looking at long time spans of hundreds of years. The method is designed to identify stages of linguistic change in general language. Our problem requires looking at the language of different groups of individuals over a much shorter timespan – looking at months over a few years, rather than decades over centuries. We will investigate the usefulness of this method in these very different circumstances.

4.5.2 Applying VNC to Hansard

We implemented a version of VNC and used it to cluster the windows of posts from our dataset described in Section 4.2. Initially, windows of 15,000 posts were used, with a step of 15,000. Each window was represented by a vector containing the relative frequency for each word in our feature set. We experimented with two different feature sets: a simple Bag-of-Words looking at the 1000 most common words, and a feature set of Brexit keywords (see Section 4.3.5). The keywords were words that were overused in Brexit-related texts compared to the rest of the corpus. Cosine distance was employed as a similarity metric, due to it being widely used to compare text similarities in natural language processing. This process can be applied to any number of groups, but for this section we compared Conservative and Labour. For easier comparison, we plotted the dendrograms horizontally rather than vertically.

Figure 4.15 shows a VNC comparison of Conservative and Labour MPs. Firstly, it is worth mentioning that the group with more data (Conservatives) forms clusters at lower cosine distances. This either suggests that this group is generally more consistent in its language, or that data quantity has an effect. Some clusters seem to have emerged, for example from the window starting in November 2015, to the window starting in July 2017. Despite there being some apparent clusters, the plot is generally very confusing and difficult to read. For example, the dendrogram occasionally goes in the wrong direction, with parent clusters clustering at lower cosine distances than their children. This suggests that neighbouring windows are less similar than more distant ones. It is possible that the window size is too small, or that the BoW features do not cluster well. Another possible problem is that we are looking at all contributions, covering a wide breadth of topics. Looking at a more focused corpus might yield better results.

To further test the method, and get clearer clusters, we next experimented with larger windows. The problem with large windows is that they usually lead to fewer windows, which would make it difficult to find stages in the data. Overlapping windows were used as the solution to this problem. This may have significant effects on VNC, as each window will have a significant overlap with its neighbours, making them much more similar to each other. However, as this is true of all windows, the neighbour which a cluster pairs with is still meaningful.

Figure 4.16 shows the VNC dendrograms of Conservative and Labour with a

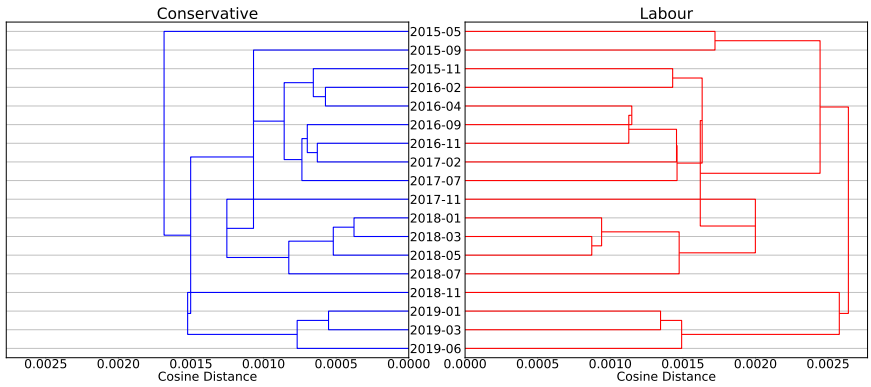


Figure 4.15: Dendrograms created using VNC for Conservative (Blue) and Labour (Red) MPs. Window size and step 15,000. 1000 most common words used as features.

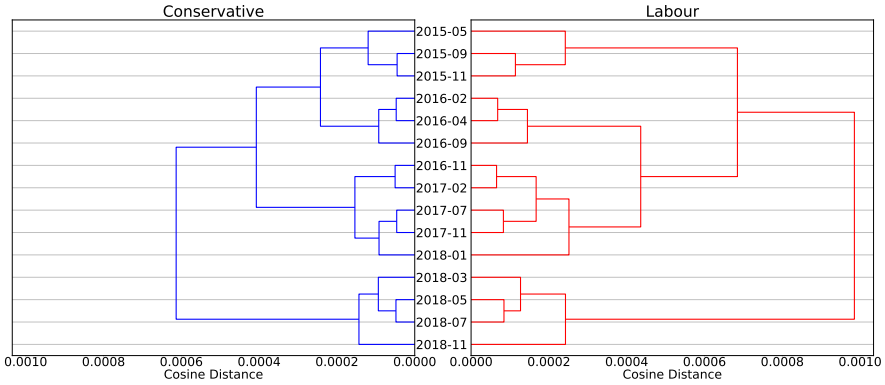


Figure 4.16: Dendrograms created using VNC for Conservative (Blue) and Labour (Red) MPs. Window size 60,000 and step 15,000. 1000 most common words used as features.

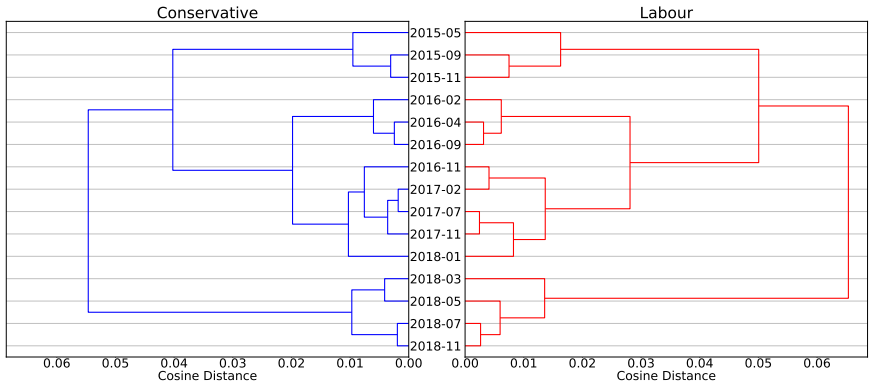


Figure 4.17: Dendrograms created using VNC for Conservative (Blue) and Labour (Red) MPs. Window size 60,000 and step 15,000. Brexit keywords used as features.

window size of 60,000. The step and features are the same as in Figure 4.15. Immediately this is more readable than the previous dendrogram. Having fewer points on the y-axis makes it much easier to distinguish clusters. The trade off is that one will miss groupings that might theoretically occur at lower levels. Trying multiple window sizes is probably the best way to navigate this problem. A balance must be met between having enough data in each window, and having sufficiently granular windows to see patterns over a short time period.

The clusters are much clearer than before. Both Conservatives and Labour appear to have four main clusters, which are the same for both parties. This could suggest that the stages of language change are shared. There is one slight difference: for Labour the two central clusters are grouped together, while for the Conservatives it is the first two. It is hard to tell, however, the reason for this difference, particularly as these clusters are merged at substantially higher cosine differences. This might suggest that it is not especially meaningful.

It is all well and good to see possible epochs in the data. But it does not really tell us anything without further investigation of which keywords are defining each cluster. To do this, we looked at the key features of each cluster to see if they follow a logical pattern. This can help us determine if the method is working, as well as possibly seeing differences between the groups.

To define clusters, we used a cut-off distance of 0.0003. In this case, this results in three stages for the Conservatives, and four for Labour. This number was arrived at relatively arbitrarily. There may be more quantitative ways of choosing the number of clusters, but we view this method more as a guide to analysis than a definitive method for finding the "true" number of stages in a corpus. Because of the clear differences caused by data size between Labour and Conservative, this number should also arguably be chosen separately for each group as it may not be comparable. The value could also be adjusted lower if one was more interested in very local clustering behaviour, or higher if one wanted to know more about higher level groupings.

We found keywords by calculating the log ratio of each word in each cluster, compared to the other clusters combined. The words with the highest log-ratios were overused in the given cluster. This will allow us to identify words which "defined" the language of that epoch. Only the words used as features were considered, but

theoretically this restriction could have been removed to get less common words that characterise each stage.

Looking at the keywords of the clusters in Figure 4.16, we found that both parties had similar keywords for each window. This makes some sense as any topic introduced by one party is likely to be discussed by the other also. Keywords early on (the first epoch for Conservative, and first two for Labour) are not EU related. Words such as "Syria" and "devolution" are highlighted as key. The first of these words relates to the war in Syria, and a significant vote about the UK carrying out airstrikes in the country. The second may relate to 2015's Scottish independence referendum. In the second Conservative epoch, and Labour's third, EU words begin to come into the conversation. Words such as "customs", "negotiations" appear as key, pointing towards the increasing domination of Brexit negotiations as a topic of debate. In the final epoch, Brexit terms are dominant. "Withdrawal", "Brexit", "referendum", "voted" are all key. These results paint a picture of both parties gradually discussing Brexit more and more throughout the period we are looking at.

By analysing these keywords, we can better understand the language that defines stages in our corpus, and whether they differ between groups. The examples given have been very simple, and designed to show how the method works. They suggest that this type of analysis could be very useful for investigating stages in language, even over short time-spans. Deeper analysis from corpus linguists and political scientists could help reveal interesting characteristics of these groups. Experimenting with different features may also reveal deeper change than what we have shown, which mostly suggests dominating topics of discussion rather than any low level language change.

One other feature set we did experiment with was Brexit keywords. These were words that were overused in Brexit-related contributions within our corpus. Figure 4.17 shows the VNC dendrogram for this feature set. The epochs are exactly the same as they were with BoW, although the distances at which windows were clustered is much greater. The key features for these epochs are more Brexit-specific, however. We used a cut-off of 0.03, which created three epochs for each group. The first epoch contained some Europe related terms, but notably featured "renegotiation" as a key term. This suggests that at this point, the discussion around the UK's membership of the EU involved renegotiation rather than leaving entirely. The second epoch has keywords

relating to specific details of the UK's departure, such as "Euratom" (European Atomic Energy Community). Finally, the third epoch contains key terms such as "backstop" - the important questions that needed to be resolved before the deal was voted on. This gives us some impression of how the topics of Brexit discussion evolved in parliament, and suggests epochs of this development.

4.5.3 Concluding Remarks

In this section we have explained the use of the VNC method for comparing groups over a relatively short time range. Although the method is still slightly opaque, we went some way to explaining the epochs using keyword analysis. Our results may have been improved by using smaller, more tailored feature sets that aimed to look at a specific aspect of language change, for example, stylistic change.

The results in comparing groups were mixed. There was an inherent problem that our techniques mainly picked up topical features, and in a parliamentary setting both parties will tend to bring up the same topics. VNC does not seem to be fine-grained enough for answering questions about who is leading discussion, etc, but it still gives us a way of splitting a corpus into epochs. Data quantity also seemed to affect the comparability of groups of different sizes. Despite this, the method can still be used, even if it is more of a qualitative guide, or support method.

We believe that VNC is a useful method, that can be used for the purpose of community analysis. It may not tell us everything by itself, but as part of a wider suite of tools it has an important use.

4.6 Fluctuation Analysis

In this section, we will discuss two methods, UFA and KFA, and their applicability for comparing the language of groups over time in Hansard. These methods aim to visualise slightly different things. **Usage Fluctuation Analysis (UFA)** is a technique for observing the way that a word's collocates, and by extension its usage, change over time. **Keyword Fluctuation Analysis (KFA)**, meanwhile, is a method for visualising the more general language change of an entire group based on their keywords for a given topic. We will provide examples of how these methods can be used for analysing the

language of Brexit over time in the Hansard corpus.

In Section 4.4, we have already identified the words with the most substantial changes in their meaning. UFA provides a way to visualise this change on a more granular level, which may be able to assist in analysis. We will also discuss the suitability of the method for comparing groups of contributors.

4.6.1 Background

Gabrielatos et al. [2012] introduced a technique called Peaks and Troughs for examining the change of a linguistic feature over time. The technique is essentially applying a GAM (Generalised Additive Model) non-linear regression [Wood, 2017] to the relative frequencies of a given word over time. By doing this, the development of a linguistic variable can be tracked all the way through a diachronic corpus. Identifying deviations from the norm can help find interesting changes in the frequency of a given feature.

Usage Fluctuation Analyis (UFA) [McEnery et al., 2019] is an extension of this technique which looks at how the collocates of a word change over time. This can tell us a lot about the changing usage of a word throughout history, for example. The first step of UFA is finding a list of collocates for a given word, at multiple, overlapping time windows. These collocates are represented by a binary vector with a 1 for each collocate which is present in a given window, and a 0 for all those that are not present.

The next step is calculating the similarity of the collocate vectors between each time window and the next. McEnery et al. [2019] used the AC1 agreement statistic [Gwet, 2008], but theoretically any appropriate similarity measure can be used. Once we have these values, the GAM non-linear regression can be applied to produce a smooth curve through the data points. Any non-linear regression could be used, but we chose to use GAM because it is what was used in both of the mentioned previous works [Gabrielatos et al., 2012, McEnery et al., 2019].

Finally, the Peaks and Troughs technique is applied to these similarities to show how they fluctuate over time. From this graph you can see when a word's usage changes, or becomes stable. This can be followed by a more qualitative analysis to better understand why the word usage changes.

4.6.2 UFA on Hansard

The first thing we did was implement a version of the UFA method in python. For finding collocates we looked at 5 word windows on either side of each word. A mutual information cut-off of 3 was used for selecting collocates, as was used by McEnery et al. [2019]. For our rolling windows, we used a window size of 10,000 and step of 2,000, looking at only EU mentions. Contribution windows were used so each window contained similar amounts of text, and only EU mentions were considered as we were interested in how words changed in the context of Brexit.

Figure 4.18 demonstrates the use of UFA on some examples of highly changing words discussed in Section 4.4. These lines can be difficult to understand intuitively, but in simple terms, we are interested in whether they are going up or down. When a word's line goes up, its usage is stabilising and when it goes down, it is changing its usage, losing or gaining new collocates. Peaks represent points of stability in usage, and troughs represent points of instability.

We can immediately see some interesting behaviours. The words "Brexit" and "customs" gradually descend, suggesting that they are changing in their usage. "Brexit", for example begins with 75 collocates in the first window, and by the final window has 406. By looking at the most significant collocates for each window, we can infer the reasoning behind the changes. Early on in the time range, "Brexit" collocated strongly with words such as "soft", and "post". Later on, words such as "softer", "cliff", "blindfold", and "crash" ranked highly as collocates, which potentially tells you a lot about how it went.

"Single" and "EU" have the opposite behaviour. Over time, "single" becomes more stable in its usage, converging towards mostly being used to discuss the single market. The word "deal" is stable in how much its usage fluctuates until late 2017, when it begins to dip and become much less stable in its usage.

We also used the method for comparing groups. To do this, we created a collocate vector at each window, for both Labour and Conservative members. Figure 4.19 shows an example of the fluctuation of Labour and Conservative MPs being compared for the words "Brexit" and "EU". In most cases the groups fluctuate in a very similar manner, as demonstrated by the plot for "Brexit". Because the method looks at presence or absence of collocates, this is not necessarily surprising as any word mentioned by one party

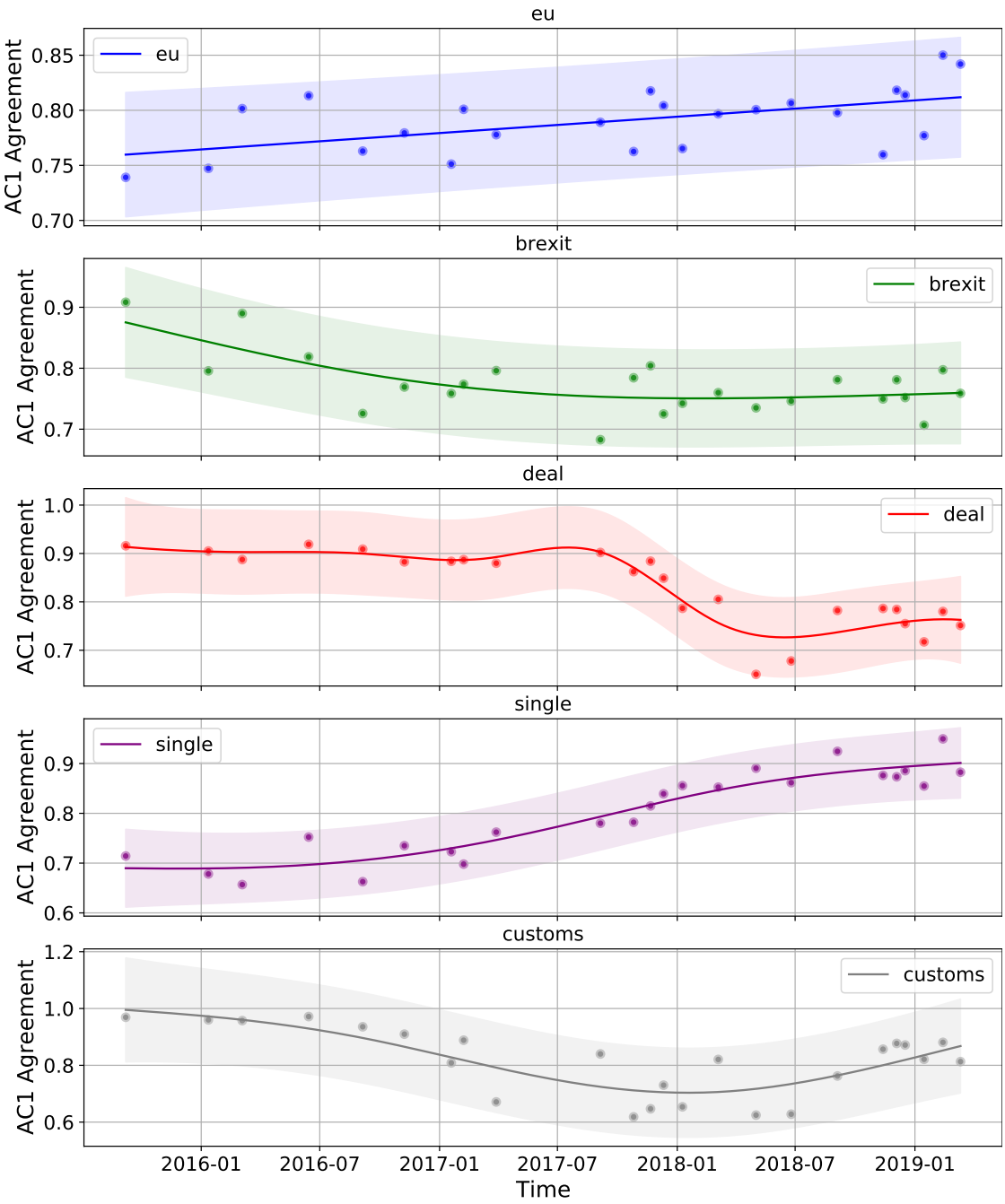


Figure 4.18: An example of UFA for five of the most changing words from Section 4.4.

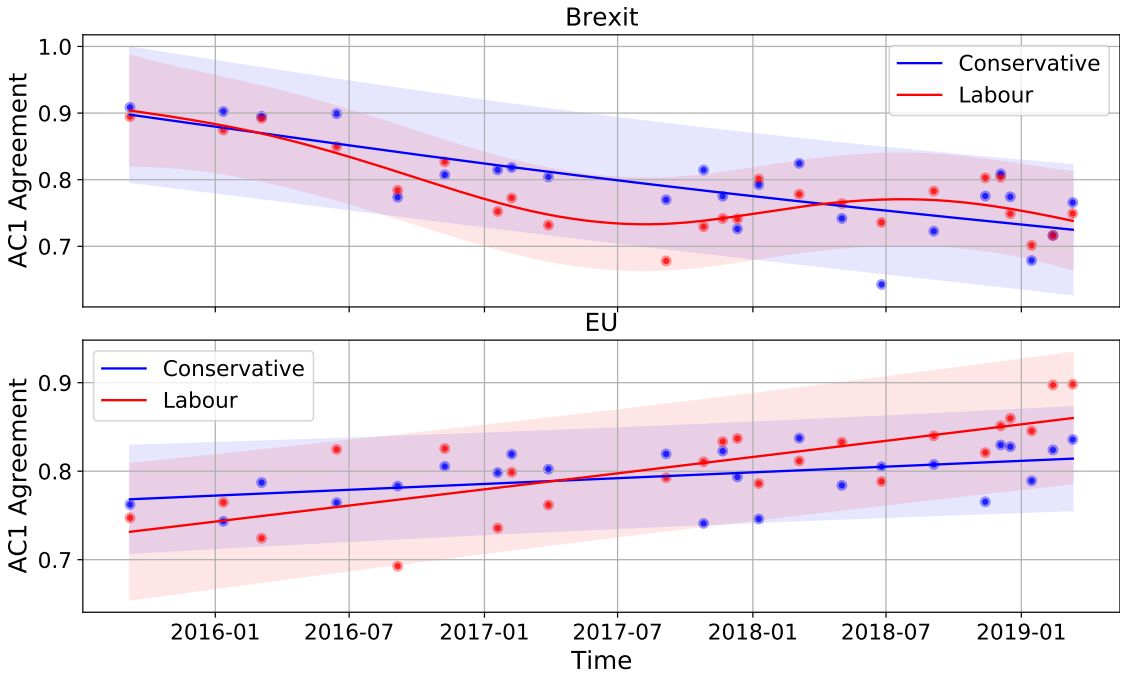


Figure 4.19: A demonstration of a group comparison made using UFA.

is likely to be used by the other in response. This highlights a key problem with the method for group comparisons, at least in a debate setting. There are some differences, as shown in the "EU" subplot. Here we see that Labour becomes more stable in its usage of the term than the Conservatives.

Looking at the top ranking collocates in each window can help to explain the differences between groups. For example, the highest ranked collocate of "single" for the Conservatives is "incompatible", while for Labour words such as "staying", "retaining", and "access" are highly ranked. This highlights the difference in attitude between the two groups, even when their collocate vectors would be very similar because of their binary nature. It is possible that UFA is simply not a method well suited to comparing groups in a debate setting, where groups use many of the same words. In other settings it may be more useful for this purpose.

4.6.3 Keyword Fluctuation Analysis

With UFA we can visualise how a word's usage changes over time. However, we would also like to see a more general representation of how different groups change in their language, relative to one another, over time. For this, we will use Keyword Fluctuation Analysis (KFA), an adaptation of the UFA/Peaks and Troughs methods [Gabrielatos

et al., 2012, McEnery et al., 2019]. Instead of finding the collocates for each window, we create a vector of keywords for a given topic (Brexit, in our case), for each group. We recorded the frequency of each keyword in each window, for EU-related contributions. The window size was 50,000, and step 10,000. Unlike with UFA, we looked at the entire corpus as we needed to find keywords for the EU related contributions with the non-EU contributions as a reference. For calculating the similarities, we used the AC1 agreement statistic, as we did for UFA. By plotting the similarities of the groups' keywords across time and between groups we can show:

1. How consistent a group's keyword usage is over time.

2. How groups change relative to one another, with respect to their keyword usage.

4.6.4 KFA on Hansard

To demonstrate its use, we used KFA to compare the language of Labour and Conservative MPs over time. We used the same hyperparameters as we did for UFA, except finding for each group instead of collocates. Keywords were considered to be words that were more prevalent in Brexit-related contributions with log-ratios greater than one and frequency greater than 10 in a given window.

Figure 4.20 shows a plot of the fluctuation of the Conservative and Labour groups over time. The plot for both groups suggests a general decrease of consistency over time. After 2017, the decrease levels out, with a slight peak in 2018. Both groups follow a very similar trajectory, even if Labour does start more consistent, suggesting that Brexit discourse in general became more varied over time, and one party's argumentation was not significantly less or more varied than the other's.

Figure 4.21 shows the KFA comparison between Labour and the Conservatives at each window. The graph shows a general decrease in agreement between the groups over time, suggesting that they became less similar in their Brexit language. This divergence is almost identical in its shape to the fluctuation plot. While it is possible it is a coincidence, plotting the comparison plot alongside the summed frequency of the keywords in each window (Figure 4.22) suggests that the agreement is essentially the inverse of the number of keywords. The more keywords there are, the less stable the keyword vectors are between groups and windows. As Brexit becomes a more popular

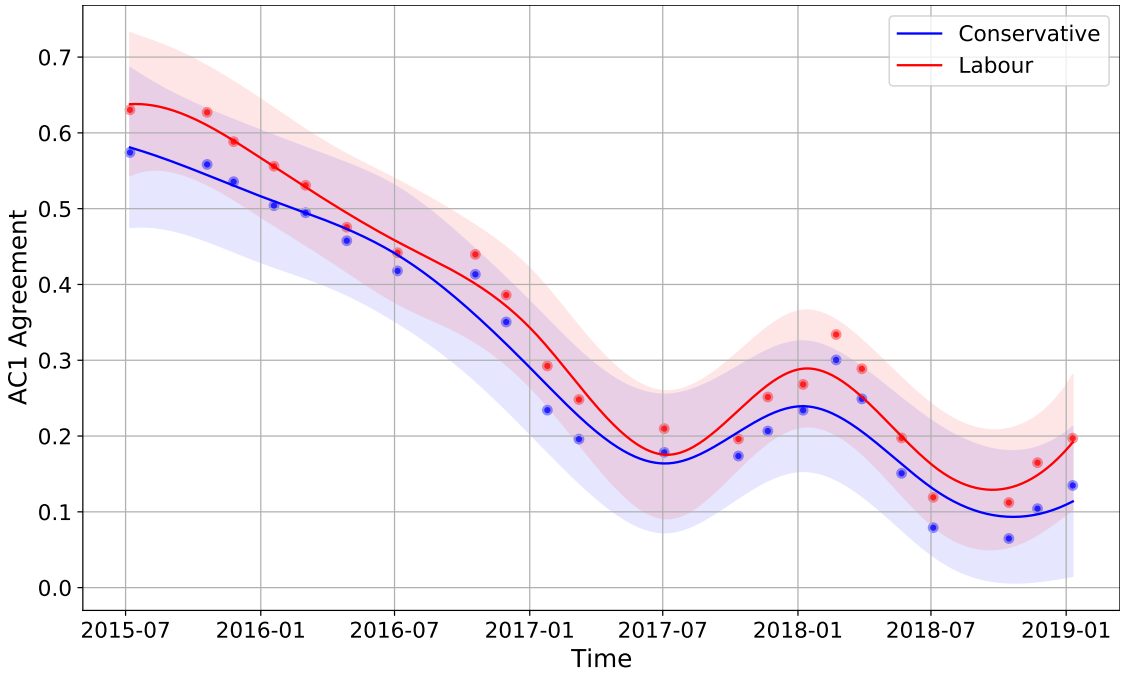


Figure 4.20: A fluctuation plot using KFA for Labour and Conservative groups, showing fluctuation of each group's keywords over time.

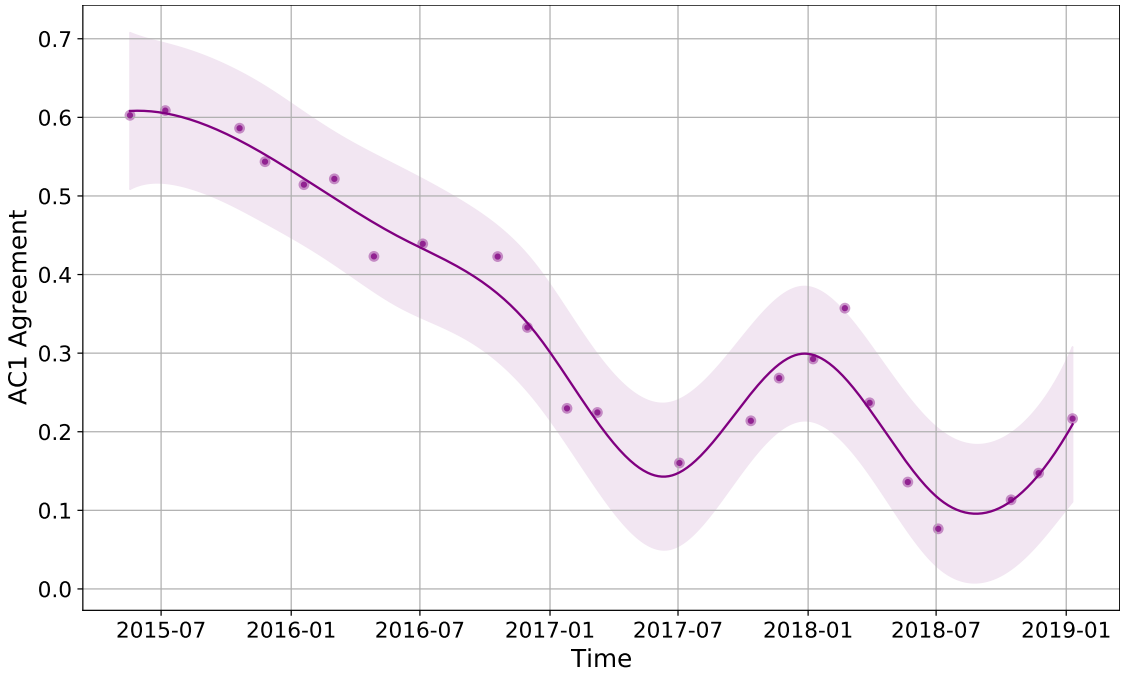


Figure 4.21: A demonstration of a group comparison made using KFA, showing a comparison of Labour and Conservative keywords at each window.

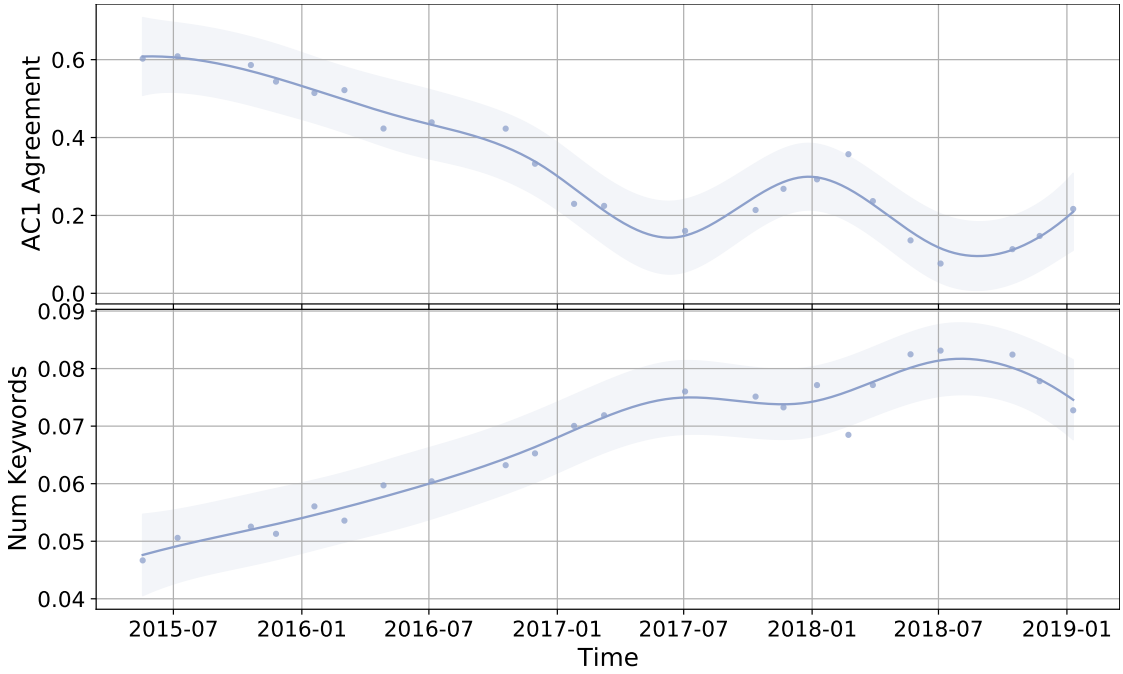


Figure 4.22: KFA group comparison, plotted alongside the total number of Brexit keywords per window.

topic of debate, it gains a wider variety of keywords and discussion diversifies. This may be a perfectly interesting finding, but it did not require this technique to uncover.

4.6.5 Concluding Remarks

In this section we have shown how to visualise language change over time by plotting the fluctuation of collocates (UFA) and keywords (KFA). These methods were applied to examine UK Parliamentary Brexit discourse between 2015 and 2019. We found that UFA had value in observing the usage change of words in general. However, it was not particularly useful for comparing groups. This was because both groups tended to use the same words in a debate setting, meaning there was little difference between the fluctuation of groups.

We also introduced the KFA method, for comparing the keyword usage of groups over time. While it did potentially highlight areas of interest to further investigate, it was limited in its usefulness. The fluctuation was overly influenced by the increasing prevalence of Brexit as a topic over time. Both UFA and KFA may be more effective in settings where the volume of discussion is more consistent, because then the differences would be caused by changes in collocates/keywords, rather than a greater variety.

4.7 Conclusion

In this chapter, we have produced a toolbox of methods for looking at language change in community settings. We have described each method's suitability for analysing diachronic change over relatively short time periods, and for comparing sub-groups over time. The various benefits and limitations of these methods were also discussed.

Overall, we found that each method was suitable under different circumstances. VNC (Section 4.5), for instance, is an effective tool for breaking a corpus down into time periods, but it is not granular enough to identify particular events. Word vectors (Section 4.4) can be used to identify words and topics that change substantially, but require a lot of data to produce results without distracting noise. Once these words and topics have been established, techniques such as UFA (Section 4.6) can be employed to observe the patterns of change for specific terms, although we found this method was not best suited to words which changed substantially in popularity over time.

Throughout this chapter, we have discussed the limitations of each method regarding the comparison of sub-groups within a community. Bearing these in mind, Chapter 5 will describe a novel method specifically tailored to the specific task. This new technique will be usable alongside the methods from this chapter, to thoroughly explore the way the language of groups changes, relative to other groups.

So far, we have only tested these methods on one dataset, so the next step going forward is to apply them to online communities, which will be done in Chapter 7. By looking at the way language changes in communities that share false information, we hope to learn more about the individuals in these groups. Techniques from this chapter may be useful for identifying interesting trends, and possibly even groups of users who behave similarly.

Chapter 5

A Novel Method for Comparing the Language Change of Groups

5.1 Introduction

In this chapter, we will describe a novel method for comparing the language of groups over time. Average Cross-Entropy (ACE), which is described in detail in Section 5.4, is an extension of cross-entropy, tailored towards comparing the language of sub-groups within communities. It is similar to fluctuation analysis, from Section 4.6, only using a different technique to track change. One main difference is that it does not focus on plotting keyword change, but rather change of a group's language model relative to another trained on a different group. This method will be an additional tool in the toolbox of language change methods we created in Chapter 4, but has been separated out into a new chapter because the novel method stands as a contribution in its own right. As with the other methods, it will help address RQ2 from Section 1.3, providing a way to study the language of sub-groups within communities over short time-spans.

This section has the following key contributions:

1. **A new method for visualising language change of groups over time.** This new method, presented in Section 5.4, uses cross-entropy with repeated sampling to visualise and compare **language unpredictability** and **language change between groups of people**, over relatively short time periods, and with limited data.

2. **An analysis of parliamentary discourse.** The method is demonstrated with a

series of case studies comparing the language of groups across political divisions in our Hansard dataset (Section 4.2). Section 5.6 investigates key questions around the Brexit debates, showing supporting correlations with known events and trends, as well as providing new insights and points of interest for further investigation.

This new approach to analysing language change of groups over a short period could be used to better understand defections, splits, and the formation of new groups. Particularly, the methods could be used to analyse other communities, such as online forums, where they could help with tasks such as finding trolls or user segmentation. This could make it a useful tool for looking at the way different groups within false information communities use language. These directions will be discussed in Section 5.8.

All code from this chapter has been made publicly available[1,2].

5.2 Cross-Entropy

Cross-entropy (CE) is a measure from information theory which allows us to compare a predicted probability distribution, Q, to a true probability distribution, P. The value of cross-entropy tells us how many bits we would need to encode an event in Q if the event was drawn from P.

The formula for calculating CE is as follows:

$$H(P, Q) = \sum_{x \in X} -P(x) \log_2 Q(x)$$

In the context of NLP, we can use this measure to compare two language models, which we consider as the probability distributions. If we use language models that obey the Markov assumption[3] [Markov, 1954], we can simplify the formula to:

$$H(P, Q) \approx \frac{1}{n} \sum_{i} \log_2 Q(w_i)$$

[1] https://github.com/dearden/language_change_methods
[2] https://github.com/dearden/thesis_language_change
[3] https://en.wikipedia.org/wiki/Markov_property

Using this formula one can calculate the CE of word sequences from a text (essentially drawn from P), according to a language model Q. We can use this to gauge how much the text we feed in deviates from this model. A higher CE suggests that a sequence is more surprising, and therefore more highly deviant, to our model. By plotting the CE of texts over time to a language model (or multiple), we can see how they converge and diverge. Cross-entropy is not symmetric, so $H(P, Q)$ is not the same as $H(Q, P)$. Kullback-Leibler Divergence (KLD) [Kullback and Leibler, 1951] (also called KL-divergence or relative entropy) is another highly related measure that can be used to compare two distributions, though it will not be used in this thesis.

5.3 Previous Work

Cross-Entropy (CE) and KL-Divergence (KLD) are strongly related methods from the field of information theory that can be used for comparing texts. Both values tell you how surprising it is that texts were generated by a given language model. Various works have calculated CE or KLD at multiple time intervals as a way of plotting language change. For example, Barron et al. [2018] used KLD to measure the novelty of speeches during the French Revolution. Tan and Lee [2015] looked at the behaviour of "wandering" users in multi-community environments (e.g. Reddit) and showed that they could predict if a user would leave a community based on their first 50 posts.

Of particular relevance is the work of Danescu-Niculescu-Mizil et al. [2013], who used cross-entropy to measure the convergence of users' language to that of two online beer communities, over their active life cycle on the forums. They achieved this by creating bigram models, which they referred to as 'Snapshot Models', for each month on the forum. By plotting the CE of posts against these snapshots, they showed how users changed relative to the forum over time.

Some related works have compared political groups over time using methods other than CE and KLD. For example, Peterson and Spirling [2018] used classification accuracy as a measure of polarisation in parliamentary discourse, and Hofmann et al. [2020] performed time series analysis with Generalised Additive Models (GAMs) to compare parliamentary speeches over time across Austrian political parties. None of these works have, however, attempted to mitigate the variable influence of specific

individuals within these groups.

5.4 Method

Our method allows the comparison of multiple groups to each other over time. As in the work by Danescu-Niculescu-Mizil et al. [2013], the method requires Snapshot models to be trained at multiple time windows. These snapshots represent the language of a group at a given point in time. The process is as follows:

1. Randomly split each group into testing and snapshot group-member samples.
2. At each window:
   (a) Train a snapshot model for each group's snapshot samples.
   (b) Calculate, for each window, the CE of each group's test samples against every other group's snapshot model and record an average per group.
3. Repeat the process multiple times with different random group-member samples and calculate mean of the average CE across all runs, for each window.

This results in a series of values for each group that we will refer to as **Average Cross-Entropy (ACE)**. This can be formulated as follows, where x and y are groups, n is the number of runs, and i is the number of texts in x_{test}. For each run, x and y are split into random test and snapshot samples.

$$ACE(x, y) = \frac{1}{n} \sum_{n} \frac{1}{i} \sum_{i} H(x_{test_i}, y_{snap})$$

Rather than comparing entire groups, the method involves repeatedly splitting group-members into "Snapshot" and "Test" samples for each group. The snapshot sample is used to train the snapshots, and the test sample is used to calculate CEs according to a given snapshot. Performing the sampling multiple times means that the stability of the difference between groups can be observed by calculating the standard deviation of CEs across all runs.

Our method is flexible enough to be applied to any set of groups in which there is no member overlap. Because test and snapshot samples are kept separate, we can compare

all groups with each other, and also with themselves. Comparing a group's test samples to its own snapshot models will give an impression of how stable the language is within that group, which we will refer to as ***Unpredictability***.

Two varieties of window may be employed, based on the dataset being used. The first option is to have each window represent a specific amount of time, for example a month, and contain all texts from the time frame. This option is arguably more logical, and ensures that windows are regular over time, but favours a dataset with a consistent number of texts at each time point. Alternatively, one can have each window contain a specified number of texts, which guarantees all windows contain the same amount of data, even if there are gaps in the corpus.

The size and step of these windows, and whether or not they overlap, are hyperparameters that can be tuned in our method. Experimenting with different configurations, the method was found to be stable across many setups. These parameters, therefore, can be set in a way that enables a balance of window size and regularity for the data being used.

The choice of language model used to calculate cross entropy will dictate the language differences observed. Separate language models are needed for each group, in each window, and for every repeated sample. Given the number of models to be trained, and the small amount of text in each sample, the language models need to be simple. We present analysis using a word bigram model, with Stupid Backoff [Brants et al., 2007] for smoothing. This allows for differences in overall word usage to be directly observed, including topics, jargon, and repeated phrases.

The method has several other hyperparameters. First is the number of texts for a group-member in each window. This will determine the extent to which prominent members can define a group's language model. Second is text length, which was set to 60 words to ensure that differing lengths did not affect the average cross-entropy [4]. Third is whether or not to balance the size of snap and test samples across groups so each group is the same size. This might help rule out whether effects are caused by group imbalances. The final parameter is the number of times to repeatedly sample. As with the window size parameters, we experimented with different configurations and

[4] 60 was chosen because the majority of contributions had at least this many words. As we shall discuss in Section 5.5, one can also use text chunks rather than simply truncating all contributions to the first 60 words.

found the method to be fairly stable. The results of these experiments will be described in Section 5.5.

5.5 Method Configuration

We conducted several experiments to better understand how the method worked, and to observe the effects of changing different hyperparameters. This will allow us to assess the stability of the method, and help us make decisions about which hyperparameters to choose during analysis. For these examples, we will use the Hansard data described in Section 4.2. Each example shows the ACE of EU contributions by Labour and Conservative MPs, compared to a reference corpus of non-EU contributions. The reference contains contributions by both parties, but only ever contains MPs from the snapshot samples of each group.

Window Type

The first important parameter that we experiment with is window type. As mentioned in Section 4.2.7, we have two types of window: **time windows**, where each window is a set number of days; and **contribution windows**, where each window is a set number of contributions.

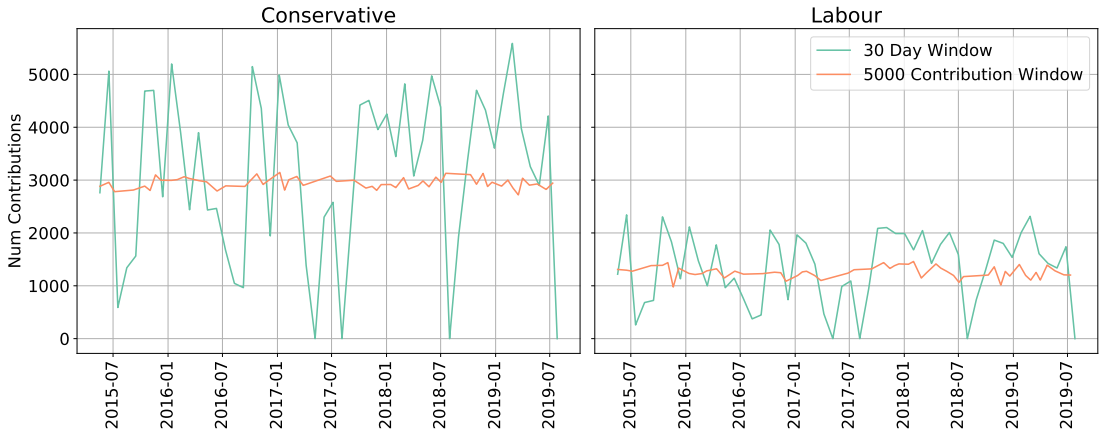


Figure 5.1: The number of contributions per group in each window across the entire corpus. The two window types are time windows, where each window is a set number of days, and contribution windows, where each window is a set number of contributions.

Figure 5.1 shows a basic meta analysis of the number of contributions per window for the two different window types. This helped us to judge which was more useful. We

found that contribution windows were more appropriate, because in this dataset there are many extended gaps in the data. For example, parliamentary recesses can mean that entire months are not accounted for in the dataset. There are also other breaks in parliamentary debate; for example leading up to an election. These holes mean that, with overlapping windows, consecutive windows may be identical, or individual windows may not have any data at all, as is the case when using small windows (e.g. a month).

Contribution windows, on the other hand, contain a consistent amount of data. The trade-off is that these windows will not be at regular intervals, so they do not relate directly to time. However, as shown in Figure 4.12 , the number of contributions over time is fairly consistent. So this may not be as much as a problem as it might seem.

The chosen window type will depend on the data being analysed, so it is important to perform at least a basic meta analysis of the number of texts over time before making a decision. In most social media datasets, for example, time windows may be completely fine because it is less likely there will be huge gaps in the data as there is in the Hansard dataset. For the following experiments, we will only look at contribution windows.

Sampling Method

We experimented with two different methods of sampling. Both methods involved sampling 60% of the speakers in each group and training snapshots on these speakers' contributions, while using the remaining 40% to calculate cross-entropy against the snapshot models. The difference between the methods is that in the first method we used all contributions for each MP, and in the second we limited each member to up to n contributions per window. Figure 5.2 shows an example of how this affects the number of contributions over time for each group.

Unsurprisingly, sampling with no limit per MP yields a larger number of contributions. The number of contributions in each window is also more variable in this setting. Imposing a limit on contributions per MP both reduces the number of contributions and also smooths out this variability. A larger value of n increases the amount of data for all groups.

These methods may show slightly different things. When a limit is in place, all MPs will essentially be weighted the same by the model, excluding those who contribute

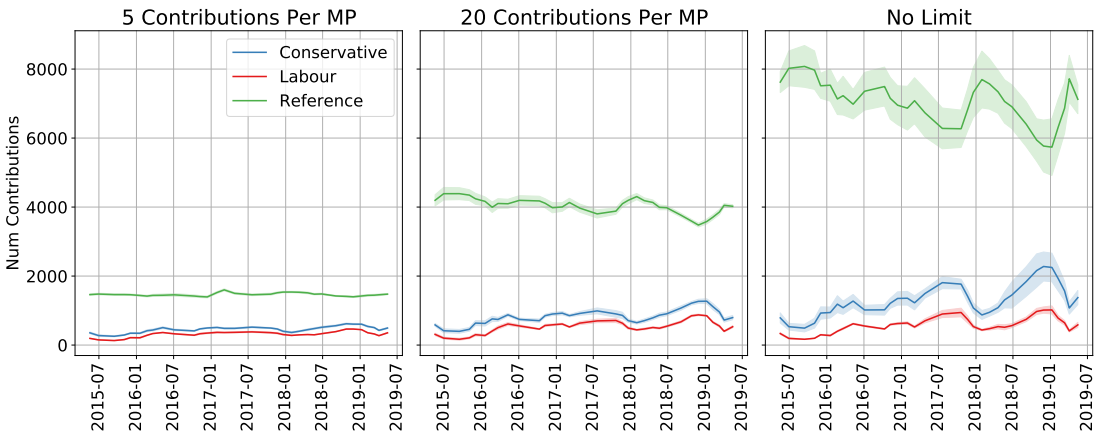


Figure 5.2: The average size across all runs of the snapshot samples per window for each group with three different types of sampling: sampling 5 and 20 contributions per MP, and not limiting the number of contributions per MP. Windows Size is 15,000 contributions and step is 5,000.

less than n times. This means that individual prolific speakers will not overly influence the language model. This may not always be a desirable trait, however. If a group contains prolific speakers, one could argue that the language model of that group should be influenced more greatly by those who speak more. For example, in a parliamentary setting, the Prime Minister and their cabinet should perhaps influence the language model more than back bench MPs. If this behaviour is desired, then it would be better to use the method with no limit. The chosen sampling method depends on what question is being answered, and how individual speakers should be weighted. By increasing n in the sampling with a limit, one could create a compromise between the two.

When looking at cross-entropies produced by the two sampling methods, as shown in Figure 5.3, cross-entropy seems lower for higher values of n (no limit being essentially n=infinity). Despite this difference, the graph's overall shape is consistent across all sampling methods. The change in n appears to simply translate the line up the y-axis. This might suggest that the factors we discussed above, to do with individual authors not overpowering the model, may not be important in this example. However, in other cases it could make a difference so it is important to bear in mind when making a decision of whether, and how much, to limit speakers.

For the remainder of this section, we will mainly use the second method, where each MP provides a limited number of contributions to each window. This is to avoid prominent MPs who speak frequently from overpowering any effects caused by MPs who speak less frequently. The method is not perfect – some MPs speak fewer than

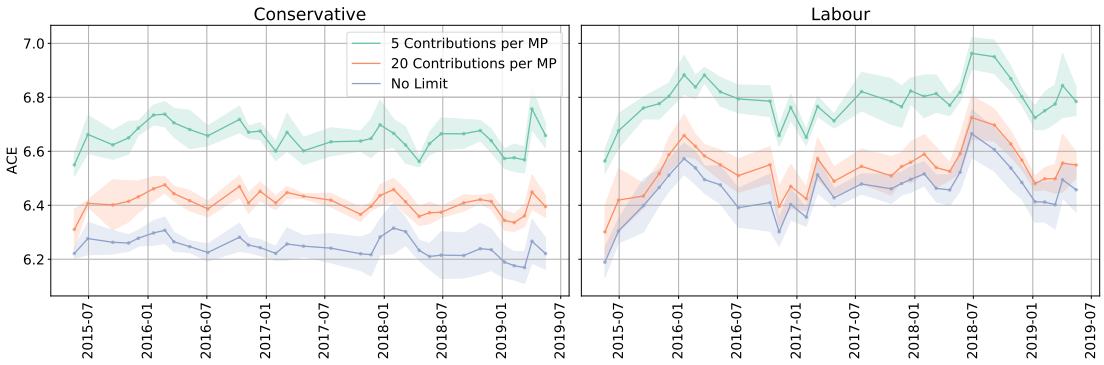


Figure 5.3: The average cross entropy across runs for each window, with three different sampling setups: sampling 5 and 20 contributions per MP, and not imposing a limit per MP. The left and right graphs show the cross-entropies for the Conservatives and Labour, respectively. These plots use a window size of 15,000 contributions and a step of 5,000 contributions.

five times in a window – but it certainly goes some way to avoid this problem. Another advantage of this method is that it does not take as long to compute, due to the limited number of contributions per window.

Balancing Groups

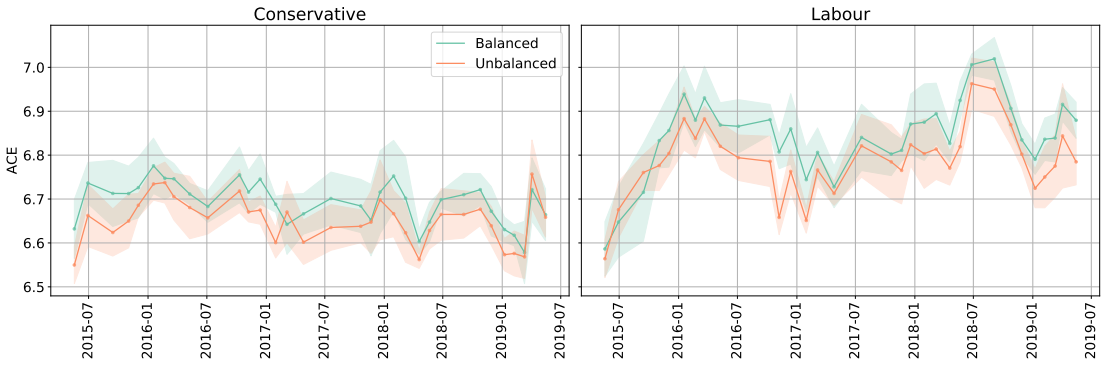


Figure 5.4: The average cross entropy across runs for each window, with and without balancing the number of MPs in each sample. The left and right graphs show the cross-entropies for the Conservatives and Labour, respectively. These plots use a window size of 15,000 contributions and a step of 5,000 contributions.

In the sampling methods explained above, the groups are not explicitly balanced, meaning that the samples of one group may contain more speakers than another. This may cause a smaller group to be more distant from the reference corpus than a larger one. To test this, we modified the sampling slightly to see whether balancing the number of speakers in each sample, so all groups have the same number, would affect the results.

In this edited method, all group samples contain the same number of speakers as the smallest group. This will not ensure that the same number of texts are in each window, but will mean that the same number of speakers are present. This, combined with the limit on contributions per speaker, will ensure that the data is far more balanced.

Figure 5.4 shows a comparison between the unbalanced and balanced samples. As one can see, the shape of both graphs are very similar for this case in which the window size is 15,000 and the step is 5,000. This suggests that, in this case at least, balancing does not seem to have a large effect on the cross-entropy over time.

Window Size

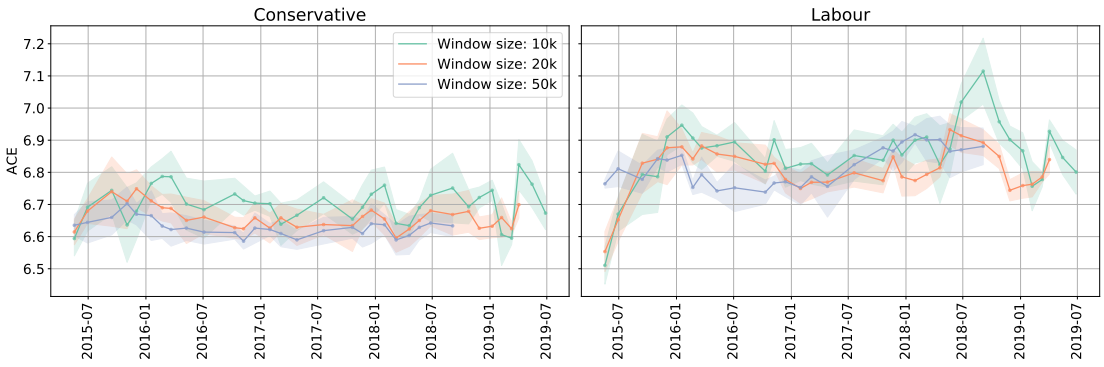


Figure 5.5: The average cross entropy across runs for each window, with three different window sizes: 10,000, 20,000, and 50,000 contributions. The left and right graphs show the cross-entropies for the Conservatives and Labour, respectively. These plots use a window step of 5,000 contributions, and they are both balanced with a limit of 5 contributions per MP.

Another important parameter of the windowing is window size. This affects how much data could potentially be in each sample. Finding the best value is a trade off between having a large number of contributions to pick from at each step, and not smoothing out and ignoring granular, short term change.

The chosen window size will depend on what is to be shown. Larger windows contain a wider time range of contributions, so can be used to look at more general change, while ignoring small, temporary changes. They may also be necessary if you are dealing with small groups, as there will be more data to sample from. Smaller windows will show more local, ephemeral changes but may be noisier because of this. Also, small windows require larger and denser data to ensure there are enough samples at each time step.

It is worth noting that, in this case, large windows do not necessarily contain more data than small windows because we are using the with-limit sampling. In practice, larger windows will contain slightly more data because more speakers will have reached the limit, however if one sampled without a limit, larger windows would contain many more contributions.

Figure 5.5 shows a comparison of three window sizes used for calculating cross-entropy. These figures seem to show the same pattern across the three window sizes. The Labour plot in particular suggests that increasing the window size flattens and smooths the shape of the line. This is what we would expect to see. Despite this flattening effect, the general shape of the line is similar across the three window sizes, with a slight peak near the beginning, and another near the end.

Window Step

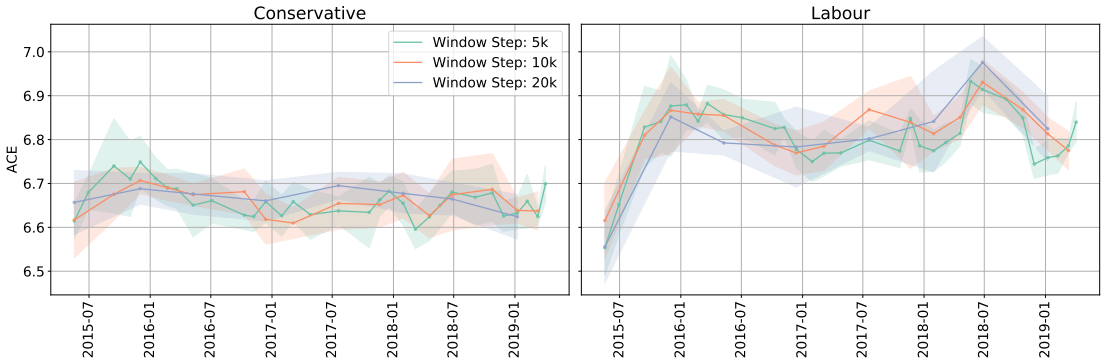


Figure 5.6: The average cross entropy across runs for each window, with three different window sizes: 5,000, 10,000, and 20,000 contributions. The left and right graphs show the cross-entropies for the Conservatives and Labour, respectively. These plots use a window size of 20,000 contributions, and they are both balanced with a limit of 5 contributions per MP.

Next, we experimented with different values of window step. We did this by keeping a consistent window size, and calculating cross entropy with windows sliding by different amounts.

A smaller step will mean more data points. This may help to identify more granular changes in the data. Steps can take any value, all the way down to a single contribution, however this would be heavy on computation, because cross-entropy would be calculated at many more time points.

The maximum step would be the window size. At this point the windows would not be overlapping. The main advantages of non-overlapping windows are simplicity and interpretability. Fewer cross-entropies need to be calculated, which leads to less computation. Also, when interpreting the results, it is easier to assign real world events to specific windows, as an event cannot take place in multiple windows.

These will be the factors one must consider when choosing a window step. Below we will compare several different steps to observe the differences. If there is little difference, it may be worth choosing the simpler, and computationally cheaper, non-overlapping window method.

Figure 5.6 shows the results of our experiments with window step. Here the lines are very similar for different steps. This is somewhat reassuring as it suggests that the method is stable. The smaller steps are more variable, as one would expect, but well within the standard deviation of the larger steps.

One interesting takeaway is that the non-overlapping windows show more or less the same change as the overlapping windows. Given that non overlapping windows are less computationally expensive, and more interpretable (which we will discuss below) it seems sensible in this case to use non-overlapping windows. This may not always be the case – for example, in datasets where we might expect rapid change. Similarly, overlapping windows will still be important for small datasets where there is not enough data to support non-overlapping windows with a small enough step to yield interesting results.

Chunking Contributions

So far, we have only used the first 60 words of each speaker's contributions in the ACE method. This is motivated by a desire to have all contributions be the same length, and is the same approach taken by Danescu-Niculescu-Mizil et al. [2013]. However, it also means excluding a large amount of data. An alternative we propose here is splitting contributions into 60 word chunks. This will mean omitting much less data.

We consider this to be a hyperparameter to be selected as there are advantages to both approaches. Truncating contributions down to the first 60 words means that the ACE can be calculated more quickly, and also does not risk over-representing speakers who make long contributions. Chunking, on the other hand, ensures that more data is

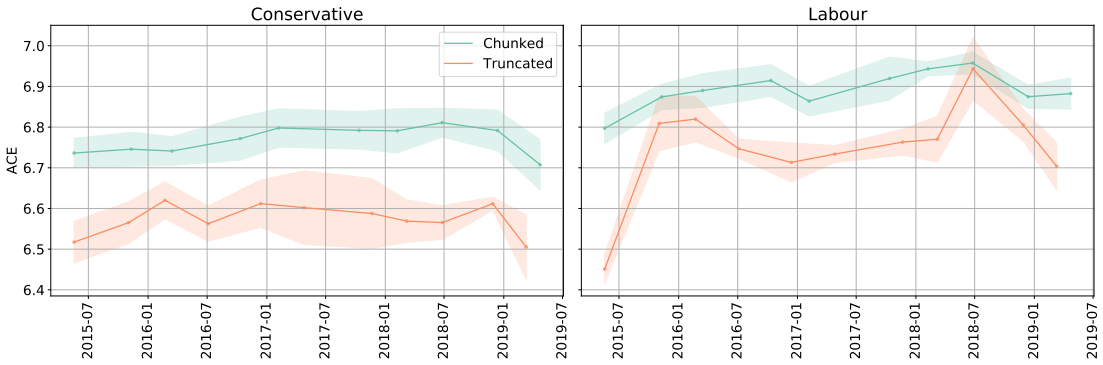


Figure 5.7: The average cross entropy across runs for each window, with two different approaches to normalising text length: truncating texts to the first 60 words, or splitting texts into 60 word chunks. The left and right graphs show the cross-entropies for the Conservatives and Labour, respectively. These plots use a window size of 15,000 contributions for truncation, and 60,000 for chunking. Both are balanced with a limits of 20 and 60 contributions per MP for truncation and chunking.

used, and also means that language from different stages of contributions is used. For example, the first 60 words is likely to contain common features such as introductions that latter chunks would not. Here we will look at the differences in the ACE plot produced by each approach.

Figure 5.7 shows a comparison of ACE plots using these two different methods for normalising text length. This hyperparameter appears to have a greater effect than the others, which is unsurprising given that it increases the amount of data substantially. This means that the distribution of speakers will be potentially different, and there is also a greater pool of posts to sample from at each window. The differences may also be affected by the fact that the windows are not identical. We tried to roughly match the number of windows by increasing the number of chunks in each window to 60,000 rather than 15,000.

Despite the differences, the graphs are not entirely different. For example, the ACE lines for Labour peak and trough at roughly the same points for both chunking and truncating. The changes in the plot are much less defined and substantial, but there is still some apparent change.

It is difficult to understand the precise reasons for the differences in these plots. Particularly because many things essentially change when this parameter is altered. The windows suddenly represent a different amount of time, and the method has substantially more data to sample from. Data size presumably has something to do

with it, but it may also be down to the inclusion of chunks other than the first. The first chunk of an MP's speech may contain common structures that are different than those of other chunks. Choosing the correct method to use may just depend on the kind of text that is being dealt with, and how representative it is desirable for the method to be of a wide range of authors.

Visualisation

A problem with the current line graphs is that they are difficult to read. This is largely because each of the "points" on the graph correspond to the beginning of a window. For overlapping windows, this means that the "new" data met for the first time at a given point can actually be met after the next point on the graph.

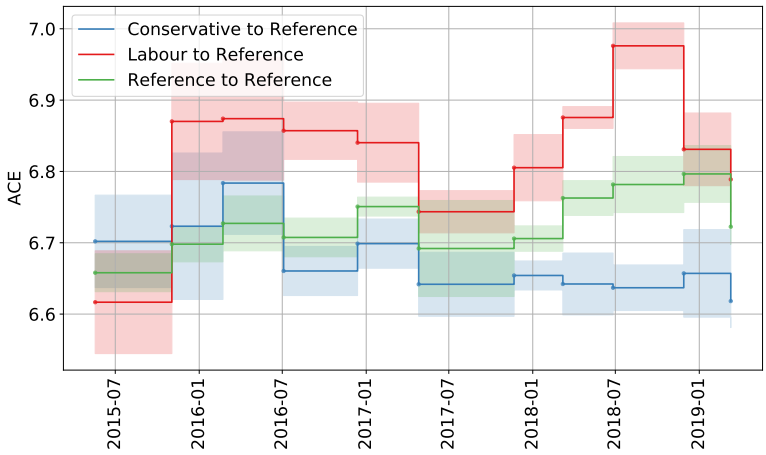


Figure 5.8: ACE of Conservative and Labour contributions to the Reference snapshot models. Unpredictability of Reference is shown alongside as a baseline to compare to. Window size and step are 15,000.

For non-overlapping windows, this problem can be fixed by producing a stepped graph, as shown in Figure 5.8. This graph makes it much clearer what the cross-entropy is for each window. However, this particular graph is only possible for non-overlapping windows. This may be fine in many situations because, as we showed above, non-overlapping windows are perfectly adequate in some cases. But we would ideally like to find a solution that works for overlapping windows as well so we can have a clear plot while maintaining the advantage of more data and smaller steps.

One possible solution is to plot line graphs using the end of the window instead of the beginning to plot each point. This is slightly easier to interpret, as each point

occurs just after the "new" data unique to that window, even for overlapping windows. Another solution might be to use an animation to display the change. The downside to this approach is that a video is not necessarily as easy to share (especially in a paper) as a static image.

The most important thing is that the person who is reading the graph knows how to read it. This user may not be aware of how the method works, and therefore it must be made very clear. For this reason, we recommend using a stepped graph wherever possible. However, if the benefits of overlapping windows are required we believe that a line graph can be sufficient, provided that the reader is fully aware of the size and step of each window.

5.6 Analysis of Hansard using ACE

To demonstrate how the method presented can be used to visualise and investigate language change amongst sub-groups, we present a series of case studies answering pertinent questions about UK parliamentary debates around Brexit.

For the following analysis we used non-overlapping windows of 15,000 contributions when looking at the entire corpus, and 3,000 when looking only at EU contributions, due to the smaller corpus size. Windows were based on contributions rather than time because there were large gaps in the data during parliamentary recesses. MPs were sampled with a 60/40 split into snapshot and testing samples. Up to 60 60-word chunks were sampled from each MP (per window), and the snapshot and training samples were balanced so that both parties had the same number of MPs in each sample. The process was repeated over 50 runs.

In the following analysis, when we say "Remainer" or "Leaver", it refers to MPs who supported Remain or Leave in the 2016 EU Referendum. MPs represent a constituency (area of the UK), and each constituency is labelled as "Leave" or "Remain", based on the vote percentage in the referendum.

5.6.1 Remainer Constituencies

The first question we would like to answer is whether Remain MPs from Leave constituencies become more similar to Leavers over time. The hypothesis here being

that MPs may change their presented stance to reflect the constituency they represent. To answer this question, we plot the ACE of two groups (Remainers from Leave and Remain constituencies) against a third group (Leavers), using the full corpus of all contributions. This is shown in Figure 5.9. Shaded areas on the graph show the standard deviation of window means across runs, indicating how variable the ACE is for each window based on sampling.

To analyse this graph, the reader can compare the lines for each group against the other. We include a third line that represents the unpredictability of Leavers – the ACE of the Leaver group against its own snapshot model. If another group appears above it, then it can be considered divergent from the Leaver language model.

Figure 5.10 plots the difference between the ACE of each group and the Leaver Unpredictability. The Leaver Unpredictability acts as a baseline, showing the increasing difference of Remainers compared to Leavers, against the Leavers language model. Significance of this difference was calculated using a two-sided student's t-test ($p < 0.01$), shown with a bold horizontal line. Significant changes between subsequent windows are indicated with a bold vertical line.

Over the timespan, both groups are significantly different from Leavers. Remainers from Remain constituencies are consistently more divergent from Leavers than those from Leave Constituencies. Even so, the difference between the divergence of both groups is steady over time, suggesting that neither group accelerates in its divergence from Leavers.

One interesting thing to note about the graph is that the unpredictability of Leavers begins to decrease after the Brexit referendum. This could suggest that the language of this group became more coherent and consistent. Figure 5.10 shows the difference between the ACE for Remainers from Leave and Remain constituencies and Leaver unpredictability. The plot shows an increase over time, particularly during the last three windows, during which many key events took place. This could suggest that it was Leavers forming a coherent group that caused divergences between the groups, rather than Remainers changing their tune.

To paint a clearer picture of how these groups diverged, we looked at EU Keywords[5]

[5]Keywords were calculated by looking at the Log-Ratio of words in the group's contributions during time period following 2018/03/15 (the final three windows where divergence increases). Only words with a frequency > 10 were considered.

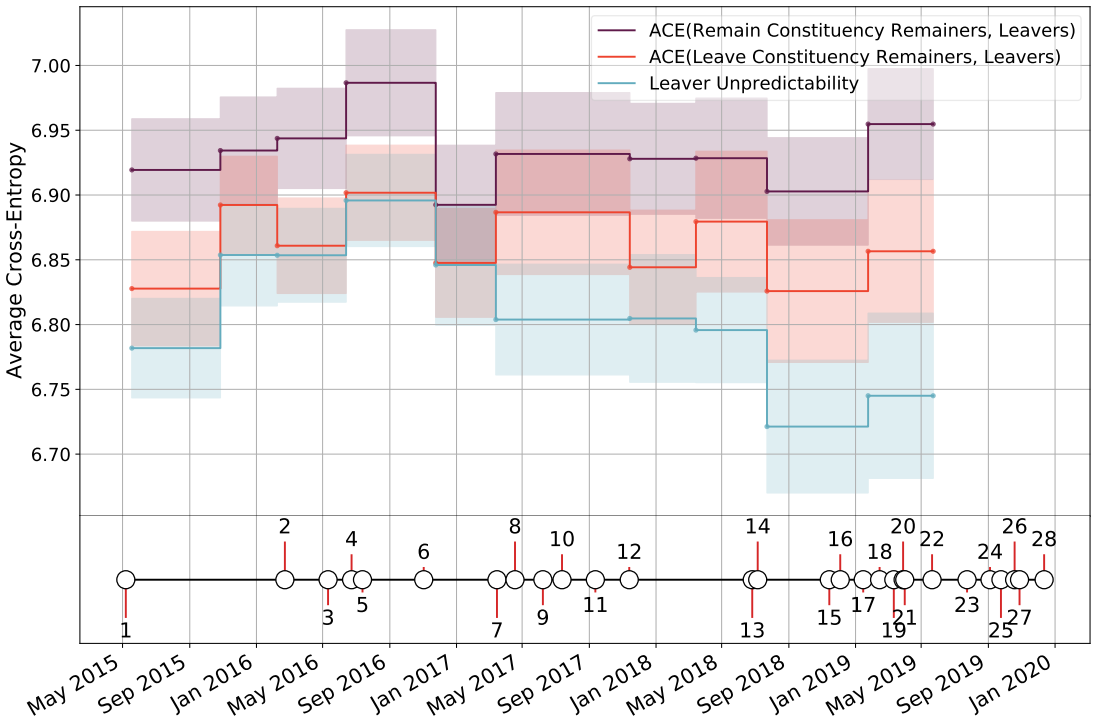


Figure 5.9: Leaver unpredictability, and ACE of Remainers from Leave and Remain constituencies against Leavers. Shading shows standard deviation across runs. Events from Figure 4.1 shown below the graph for reference.

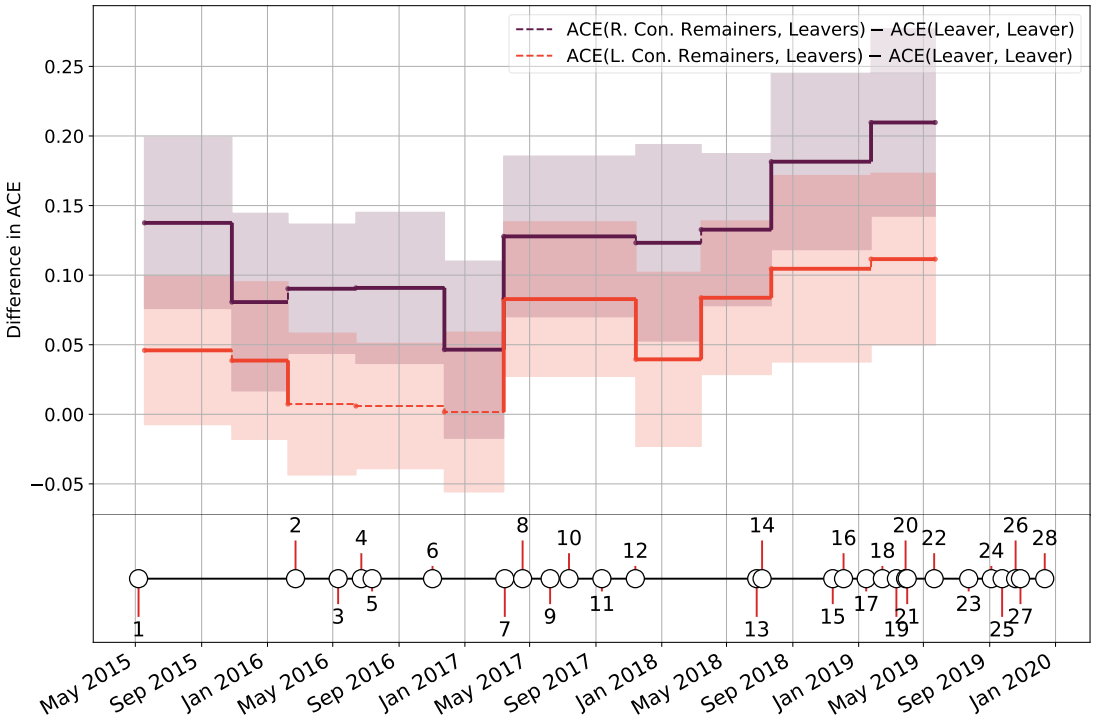
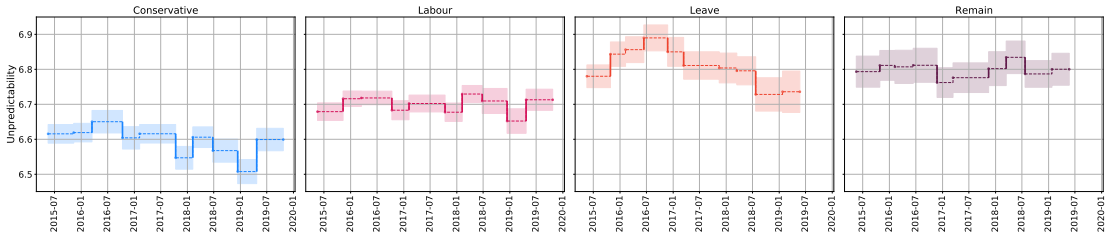
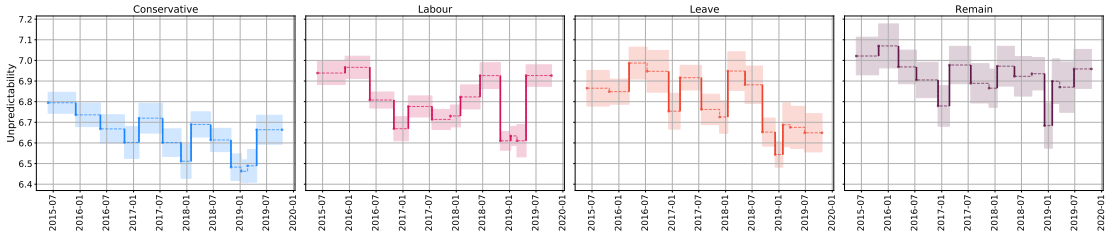


Figure 5.10: ACE of Remainers from Leave and Remain constituencies against Leavers, with Leaver unpredictability subtracted. Shading shows standard deviation across runs. Bold horizontal lines show significant differences. Bold vertical lines show significant changes between subsequent windows. Events from Figure 4.1 shown below the graph for reference.



(a) For All Contributions



(b) For EU Contributions

Figure 5.11: Unpredictability of Labour and the Conservatives over time across all contributions and EU mentions. Shading shows standard deviation across runs. Bold vertical lines are significant changes.

for each group. These were words that were overused by the group during EU-contributions compared to all non-EU contributions from the same window.

Words such as “deadlock”, “no-(deal)”, and “renegotiate” feature as words ranked highly in the Remain-Constituency Remainer group, and lower in the Leaver group. These words suggest that disagreements on the Withdrawal Agreement are key dividers between these groups. Also enlightening are the differences between Remainers from Remain and Leave constituencies. These two groups had largely very similar keywords, suggesting not too great a difference between them. The keywords of these groups compared to leavers differ in interesting ways. Words such as “hostile” and “chaotic” rank higher amongst Remain than Leave constituency Remainers. This possibly suggests greater hostility against Brexit among Remainers from Remain constituencies.

From this analysis, there is no basis to claim that Remainers from Leave constituencies converged to Leavers over time. It seems instead that Remainers stayed consistently divergent from Leavers over time, and that Remainers from Remain constituencies diverged more. Meanwhile, Leavers appear to have become more consistent in their language following the referendum.

5.6.2 Consistency of Messaging

During this period of UK politics, particularly in the run up to the General Elections in 2017 and 2019, Labour was accused of being inconsistent with its messaging around Brexit[6]. This has been suggested as a contributing factor to their defeat in both elections. To observe this characteristic with our data, we looked at the unpredictability[7] of Labour's language over time.

Figure 5.11a shows the unpredictability of Labour and the Conservatives across all contributions. Labour is generally more unpredictable in its language than the Conservatives, and both groups are relatively stable in their unpredictability. The Conservatives decrease in their unpredictability during 2018, and into 2019, though they returned to their original unpredictability for the final window. This might lend some credence to the notion that Labour was not as "on message" compared to their rivals.

Looking at EU contributions (as shown in Figure 5.11b), Labour is still more unpredictable than the Conservatives, and with much greater levels of variation. Labour is more unpredictable leading up to the referendum, but stabilise afterwards. This could suggest that Labour's message in the lead-up to the referendum was not consistent. It then later has large peaks of unpredictability during much of 2018 and 2019, with the exception of a substantial dip in late 2018/early 2019. Both parties experienced this decrease, though it is much more pronounced for Labour.

This dip corresponds to the series of indicative votes conducted on Brexit, and encompasses three windows spanning a time range from 2018/10/30 to 2019/04/24. During this period it is possible that MPs temporarily unified in terms of their messaging. Both parties have similar EU keywords, such as "Remainers" and "Brexiteers" during this period, though there are also some differences. More positive words relating to Brexit ranked higher with the Conservatives (e.g. "orderly" and "WTO[8]"), while Labour brought up certain words that suggested they wanted a softer Brexit deal (e.g. "Norway", "Erasmus"). Labour's keywords for this period compared to previous windows suggests increased questioning of details during this period (e.g. "stockpiling", "EHIC"), criticism of the Brexit process (e.g. "deadlock", "no-(deal)"),

[6] `https://tinyurl.com/labour-brex-stan`

[7] ACE of the group's test samples against its own snapshot.

[8] WTO (World Trade Organisation) Brexit was often used as a more positive spin on "no-deal Brexit".

and desire for another vote or softer Brexit (e.g. "confirmatory", "compromise", "extension").

Our findings suggest that Labour was less consistent in their language surrounding Brexit, especially during the latter stages of the process. However, MPs seemed to unify, and become less unpredictable in their language usage, during key votes.

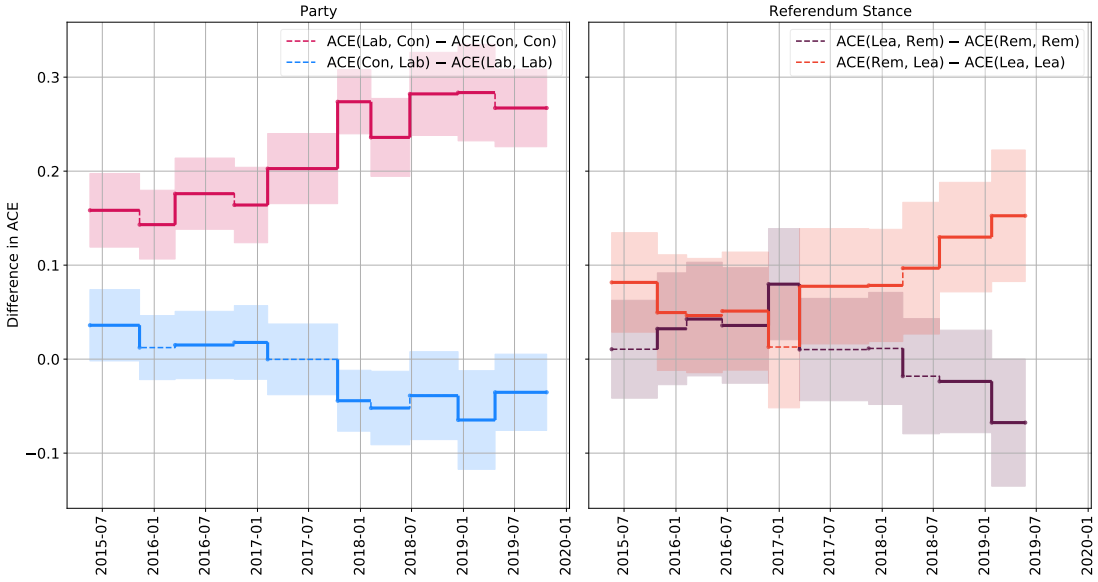
It has been suggested that the message of the Remain campaign was less clear and consistent than the romantic ideas of sovereignty pushed by Leave [Spencer and Oppermann, 2020]. We are interested if this was evident in parliamentary debate. Figure 5.11a shows the unpredictability of remain and leave supporting MPs. There does not seem to be an obvious separation as with the parties; both groups maintain a similar level of unpredictability, albeit with a notable dip in early 2019. One interesting feature of the Leave graph is that there is an increase in unpredictability leading up to the referendum, followed by a stabilisation afterwards. A significant drop in unpredictability occurs in late 2018, around the time the Withdrawal Agreement was announced. This could suggest that Leavers became united in their messaging behind Brexit as it became closer.

Looking at EU contributions (Figure 5.11b) we can see that the groups are not wildly different. Remain appears to be slightly more unpredictable, though the difference is not huge. Remain becomes less unpredictable up to 2017, and then fluctuates around a similar level. As with Labour, Remainer MPs briefly become much less unpredictable during the period of indicative votes. Leave fluctuates significantly, but most notably it decreases in unpredictability substantially for the last few windows.

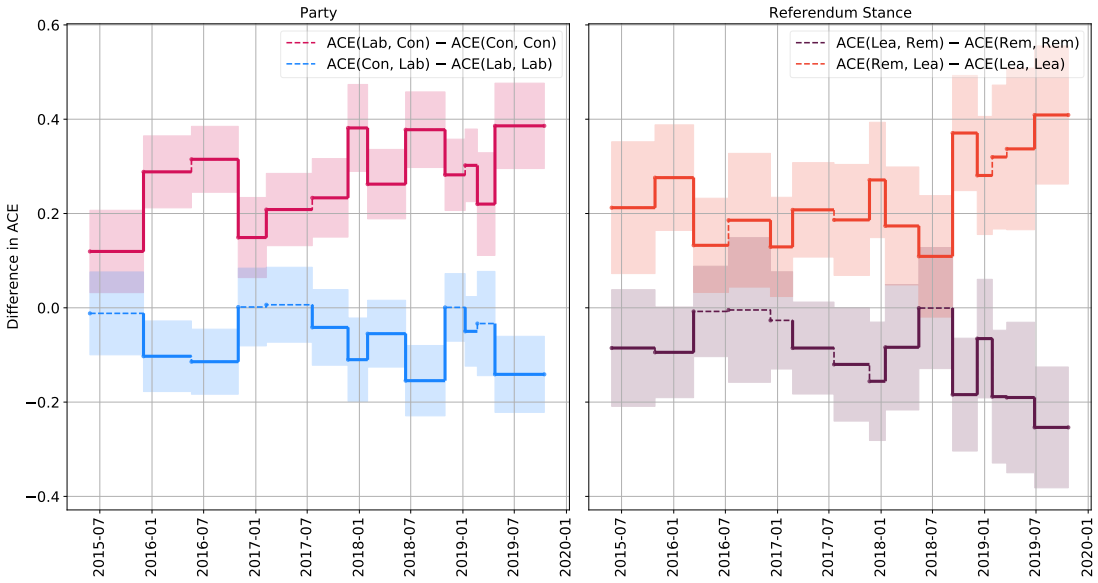
Our findings do seem to back up, to an extent, the suggestion that Remainers were less consistent in their messaging. More specifically, though, they suggest that Leavers became more unified during the period of the Brexit process where key votes occurred about details of the UK's departure, such as extensions and indicative votes.

5.6.3 Party vs Referendum Stance

Another question we were interested to answer was whether MPs are more defined by their party or their Brexit stance. To answer this with our method, we look at the difference between the cross-entropy of one group compared to another, and the unpredictability of the other group. This is conceptually similar to Figure 5.10, except



(a) All Contributions



(b) EU Contributions

Figure 5.12: Comparing groups of MPs by party and Brexit stance. Shading shows standard deviation across runs. Highlighted horizontal lines are significant differences and bold vertical lines are significant changes.

the baseline is the unpredictability of another group rather than a single group that all groups are compared to.

For example:

$$ACE(Lab, Con) - ACE(Con, Con)$$

This would tell us how much less predictable Labour is than the Conservatives, according to the Conservative language model. Another way of looking at it is that it quantifies how much worse the Conservative model is at modelling Labour contributions than Conservative contributions.

Figure 5.12a shows this for Labour and Conservative, and Leave and Remain. The fact that the differences are much greater for parties suggests that MPs are more identifiable by party than referendum stance. A similar shape of graph is produced for Leave/Remain as for Conservative/Labour, though it is dampened. This could suggest that the differences in Leave and Remain groups correlates to the Conservative to Labour differences, which would not be surprising as the vast majority of Labour MPs supported remain.

The difference between $ACE(Lab, Con)$ and Conservative unpredictability appears to increase over time. This suggest that the difference between the two parties has increased over the time span. Similarly, Remain MPs have gradually diverged from their Leave colleagues, though the divergence only begins in 2018.

If we just look at EU Contributions, we would expect the Remain/Leave groups to be more deviant from one another. Figure 5.12b shows that for parties, we get a similar increase just looking at EU contributions, though with a spike in 2016. This spike corresponds to the period surrounding the 2016 referendum.

The Leave/Remain graph is also similar, though the differences between the groups is more pronounced, as one would expect. There is a very clear jump in divergence in late 2018, after which the groups appear to be much more polarised. It suggests that MPs became more defined by their Brexit stance during the latter periods of negotiation.

The lines for both parties seem to mirror each other. This does not make sense logically, as one would think that if Labour is diverging from the Conservatives, the Conservatives would also diverge from Labour. The behaviour is caused by Labour's high unpredictability. As we have seen in Figure 5.11, Labour is more unpredictable than the Conservatives. The Conservative lines in Figure 5.12 being

around 0, suggests that the Labour model finds it no harder to model Conservative than Labour contributions. This tells us more about Labour's unpredictability than it does about the difference between Conservative and Labour. The same is true in the Leave/Remain plot. It is arguably a limitation of plotting the difference in this way, and must be considered during analysis.

From looking at these plots, we can say that Party seems to more consistently "define" the identity of MPs more than their Referendum stance. However, especially at certain time points corresponding to key events, referendum stance does appear to separate MPs. Divergences between both pairs of groups increased over time, suggesting an increase in division and polarisation across this period.

This method could potentially be used to identify events that occur in the data. For example, by looking at the peaks in divergence between the Leaver and Remainer groups, we could find significant events throughout the Brexit process. However, the resolution of the windows in the current method may not be high enough to identify granular events.

5.7 Limitations

As with any other technique, our method has its limitations. The first of these that we will discuss is the method's interpretability, or occasional lack thereof. In isolation, the value of cross-entropy does not mean very much. The number only indicates a divergence from a reference model, so two windows with identical CE values are not necessarily identical, they are merely identically divergent. Think of it like a goat attached by rope to a garden stake – the animal could roam around the stake in circles, maintaining the same distance from the stake but nevertheless always eating fresh grass. There is no "solution" to this problem other than not putting too much stake into the value of cross-entropy, and only using it as part of a wider analysis.

Another problem of interpretability is that people do not think in terms of bits and information theory, so the actual number of CE does not really mean anything to anyone. This is why we have been emphasising direct comparison (e.g. of two groups to the same reference) rather than trying to read anything into the numerical value.

The next limitation we will discuss is the simplicity of our language models. Bigram

models are very basic, and far from state of the art. It would be worth looking into using more complex models in future work, but for now bigram models suit our purposes. They are also what has been used in previous work [Danescu-Niculescu-Mizil et al., 2013]. Not to mention, there is some value in simplicity, both in terms of reduced computation and understanding the models.

The amount of data can also cause limitations. Our method is reliant on having a certain amount, the more the better. A corpus must be fairly large in order to have enough text per window. Establishing exactly how large is an interesting subject for future work. Despite this limitation, considering the abundance of data online, it seems fair to assume that for many this will not be a problem. Also, we ensured that our method made the best use of a small amount of data by using repeated sampling, etc.

One inherent limitation is in the idea of groups as a single entity. Groups are not people, and a group, for example a political party, can contain a wide spectrum of individuals and beliefs. We acknowledge this, but still think it is a worthy task as with many other author profiling tasks, where the same criticism could be made. In our case, the experiment was set up in a way such that the groups made sense for the task at hand. When designing any research using this method, it may be important to be quite specific with the groups you compare.

The final limitation that we will discuss is that the method has only been demonstrated with a very constrained example. We only looked at British English – not to mention a highly unnatural form of it – and only looked at one genre and context. This obviously needs wider investigation, but for now we are just proposing a methodology using data where we can be confident of group membership. In future work it will be important to test this method with different types of corpora from multiple languages.

5.8 Conclusion

In this chapter, we showed how cross-entropy can be used to plot language change of different groups within a community over a relatively short time span. We showed how the method could be used to compare the language of two British political parties over five years of parliamentary debate. Examples were given of graphs that plot these changes, and we demonstrated how they could be interpreted. We suggested approaches

for finding key areas, and offered a very basic example of how we hope researchers could perform a linguistic analysis guided by the results of our method.

A rigorous stability analysis was conducted of the method, showing the effect of changing different hyperparameters. We found it to be fairly stable in the environment of parliamentary debates. In future, the method will need to be tested further with different datasets. We also discussed the limitations of the method, some of which may be alleviated in future work.

Future work will apply this method to new data such as online communities, including those dedicated to false information, as we will demonstrate in Chapter 7. In such settings, there will be many more users, each contributing less data, which will be a useful test of the method's stability. Another logical extension would be to see if the approach can be adapted into an unsupervised method for highlighting potential groups of individuals within a corpus.

Having shown how it can work, we hope that our proposed method can be useful to anybody looking at the language of groups within communities, as well as those studying language change more generally. We also hope that it can provide useful insight alongside other methods for investigating diachronic corpora, be they more traditional corpus linguistic approaches, or computational analyses.

Chapter 6

An Introduction to Flat Earth Communities

6.1 Introduction

There are many communities online that spread false information, or that are dedicated to the discussion of it. Recently, work has investigated some of these communities, such as the */pol/* community on 4chan [Hine et al., 2017], or anti-vaccination discourse on the web [Kata, 2012]. Much of this work has focused on disinformation on popular social media platforms, such as Twitter and Facebook. This is largely due to their popularity, and access to data on these platforms. However, it is not always on these mainstream platforms where the disinformation originates [Zannettou et al., 2018b].

In Chapter 3, we identified problems with researching disinformation. Two of the key issues we discussed were belief and deceptive intent. To summarise in brief: whether or not somebody believes what they are saying, and whether they are trying to deceive, may dramatically change the linguistic features in any text that may be considered false information. This is especially true if treating disinformation as deception. In order to study disinformation, we need to understand how communities made up of "believers" operate, and how they use language. This may help us assess the types of people spreading disinformation, and how genuine they are.

Language change will play a key role in any analysis of online communities, as they are not static. New members come and go, and react to external events. In Chapter 4, we discussed and adapted methods for looking at language change within communities.

These methods will also be useful when studying online communities.

This chapter will focus on one fringe web community, dedicated to the discussion of false information, the Flat Earth (FE) community. The community is particularly interesting because in recent years it has received substantial attention, both online and in mainstream media [Brazil, 2020], and challenges overwhelming scientific consensus (even more so than vaccine or climate change denialism). It is also of a different topic to other types of disinformation, which makes it an interesting and novel case study. Quite understandably, research so far has largely focused on false information having substantial real world impact, such as Qanon [Papasavva et al., 2021], or anti-Vax [Kata, 2012]. But we believe that more can be learnt by looking at different sources, to build a more general impression of false information and conspiracy theories.

To our best knowledge, this is the first social media and NLP analysis of the Flat Earth community. One of the major contributions of this chapter is providing a characterisation of this community, and making observations about how it operates. This builds on previous work studying online communities [e.g. Danescu-Niculescu-Mizil et al., 2013, Zannettou et al., 2018a].

The chapter will centre around a case-study of the Flat Earth Society (FES) forum. As well as examining this forum, we will also compare it to Flat Earth subreddits (more "mainstream" Flat Earth communities), as well as to related non-FE subreddits. The Flat Earth conspiracy theory is an interesting one because it is in many ways more extreme and less believable than other conspiracy theories. Belief, whether sincere or otherwise, in something so demonstrably untrue makes for an interesting community.

Specifically we are interested in the language that defines this false information community, including the way it changes over time. We are also interested in the types of user, and styles of discussion, evident in the community, and what this suggests about the make-up of conspiracy communities, particularly in relation to belief. The analysis will seek to provide answers to the research questions from Section 1.3, particularly RQ2 and RQ3.

In Section 6.4, we perform a meta-analysis of the FES forum. This provides a general impression of the community, and will contribute to RQ3 from Section 1.3. Section 6.5 will echo this meta-analysis, but looking at Flat Earth subreddits and related non-FE subreddits for comparison. In Section 6.6, we will conduct a more

linguistic analysis of the Flat Earth community, seeing what language "defines" Flat Earth discussion, contributing to RQ1 and RQ3 from Section 1.3. All the code is available in a GitHub repository[1]. The data is also available[2].

6.2 Related Work

In this section, we will discuss the literature relating to conspiracy theories, and specifically Flat Earth Theory.

6.2.1 Conspiracy Theories

To understand the flat earth conspiracy, we need to better understand conspiracy theories more generally. Within psychology especially there has been a significant interest in conspiracy theories over the past decade [Douglas et al., 2017, van Prooijen and Douglas, 2018, Douglas et al., 2019]. In this chapter, we will use the definition of conspiracy theories used by Douglas et al. [2019]:

> "'Conspiracy Theories' are attempts to explain the ultimate causes of significant social and political events and circumstances with claims of secret plots by two or more powerful actors."

There are many famous examples of conspiracy theories. John F Kennedy's assassination spurred conspiracies about the involvement of the CIA or an accomplice to Lee Harvey Oswald [Enders and Smallpage, 2018]. The September 11th terrorist attacks in 2001 spawned theories surrounding the idea that it was staged by the Bush administration [Laine and Parakkal, 2017], and NASA has been accused of faking the moon landing [Swami et al., 2013]. Various conspiracy theories float around relating to science [Goertzel, 2010]. For example, vaccination sceptics (antivaxxers) have long spread disproven theories about the negative effects of vaccination [Jolley and Douglas, 2014b], climate change is considered a hoax by many [Jolley and Douglas, 2014a], and AIDS denialism is widespread in some groups [Hogg et al., 2017].

Such conspiracy theories can have serious real-world consequences. Between 2000 and 2005, an estimated 330,000 South Africans died because of government inaction

[1] https://github.com/dearden/thesis_flat_earth
[2] https://doi.org/10.17635/lancaster/researchdata/513

due to belief in AIDS conspiracy theories [Chigwedere et al., 2008]. Jolley and Douglas [2014b] showed that exposure to vaccine conspiracies had a negative effect on vaccination intentions. These examples show how conspiracy theories can be incredibly destructive to society, so it is vital that we better understand the way these theories are shared and the people who spread them.

Conspiracy theories are not limited to niche parts of society. Some conspiracies are very widespread. For example, Enders and Smallpage [2018] found that roughly 60% of Americans believe that the CIA murdered JFK. They are common in different cultures across the world, from traditional to modern societies [West and Sanders, 2003]. Targets for conspiracy are also varied. While many theories target governments or powerful institutions [Laine and Parakkal, 2017], others target minority groups [Kofta et al., 2005]. Conspiracy theories can also occur on a micro-level, for example within an office environment [Douglas and Leite, 2017].

Conspiracy theories have been associated with certain types of personality, and ideology. For example, Galliford and Furnham [2017] found that belief in political conspiracies was strongly correlated with right-wing political beliefs. Though it is worth noting that other works have not found this connection [Oliver and Wood, 2014], and it may be down to the chosen conspiracies or the fact that much of this work was done during the Obama administration, when anti-government thinking was predominantly right-wing. Certain demographics within society have also been found to be more prone to believing conspiracy theories [Thorburn and Bogart, 2005]. Often, it tends to be groups of people who see themselves as oppressed, and feel anxious [Grzesiak-Feldman, 2013] or powerless [Abalakina-Paap et al., 1999].

An interesting question regarding conspiracy theories is why people believe them. One concept which has been widely discussed is the idea of conspiratorial thinking, and that some people have a tendency towards it [Brotherton et al., 2013]. These people would be more prone to believe in conspiracy theories. The single greatest predictor for believing in a conspiracy theory is believing in another [Goertzel, 1994], even if they are unrelated or contradictory [Wood et al., 2012]. This suggests that a certain type of person is simply open to these kinds of ideas. Some people like to explain things by seeking patterns, even if there are none [Whitson and Galinsky, 2008].

Another prominent reason for believing in conspiracy theories is to protect already

held beliefs which are being challenged [Lewandowsky et al., 2013]. For example, if you already suspect that vaccines are bad, you may be more inclined to believe somebody who says they cause autism. Social factors can also be at play. Many people who believe conspiracy theories perceive that a powerful out-group is threatening their in-group [Imhoff and Lamberty, 2018]. Believing and sharing conspiracy theories is a way to try and protect the threatened in-group.

The features of conspiracy theory belief have been widely studied. One interesting finding is that belief in conspiracy theories is emotionally, rather than analytically, driven [Swami et al., 2014]. This seems unintuitive, as conspiracy theories often involve elaborate arguments and evidence. However, studies have found that more analytically minded people are less likely to believe in them. For example, conspiracy theory belief is less likely amongst the more highly educated, who generally tend to be more analytical [van Prooijen, 2017]. Belief in conspiracy theories involves finding patterns in random stimuli [van der Wal et al., 2018], and is often rooted in negative emotions (e.g. anxiety [Grzesiak-Feldman, 2013]).

There are other differences between believers and non-believers. Believers often make an effort to appear rational and open-minded in discussion [Wood and Douglas, 2013]. They spend a lot more time attacking the "official" explanation rather than proposing an alternative. Meanwhile, non-believers do the opposite, advocating the "official" position rather than attacking the conspiracy. Non-believers often use a more hostile tone [Golo and Galam, 2015], which can lead to believers feeling oppressed. Faasse et al. [2016] performed a comparison of the language used by anti and pro vaccination Facebook posts, finding that anti-vaccination posts were more authoritative, confident, assured, and manipulative in their language usage.

6.2.2 Flat Earth Theory

The modern flat Earth movement was arguably founded in 1956 by Samuel Shenton, with the creation of the International Flat Earth Research Society. It did not achieve widespread attention, however, until its move online in the 2000s. On the web, a "Flat Earth Society" forum was established, dedicated to the discussion of **Flat Earth Theory (FET)**. In 2013, some members of this community split off to create another forum. Over the past few years, there has been a significant increased interest in FET [Brazil,

2020], largely thanks to the prominence in flat Earth videos on YouTube[3]. There are now multiple subreddits dedicated to FET, and many videos online both advocating and debunking FET. Some well known figures have even "come out" as flat Earth believers[4], even if it was not always totally sincere.

Flat Earth Theory is largely based on the writings of Samuel Rowbotham [Rowbotham, 1865], who introduced an idea called Zetetic Astronomy, which posited the Earth as a plane, surrounded by a wall of ice. FET is often "backed up" by simple, do-it-yourself "experiments", which are favoured over traditional scientific sources. It also relies on classic conspiracist concepts such as long-running coverups conducted by powerful institutions like NASA or world governments. The mechanics of FET are not set in stone, with many different theories proposed to explain the flat Earth model. There is also, due to the nature of the belief, a link between FET and religious fundamentalism, though this is not universal amongst all FET believers [Olshansky et al., 2018].

Flat Earth Theory has barely been researched in academia, but there have been some works that looked at it over the past couple of years. Paolillo [2018] outlined some of the key characteristics and beliefs of FET, focusing especially on how it has manifested on YouTube. Landrum and Olshansky [2019] attended the Annual Flat Earth International Conference in 2017, and interviewed Flat Earthers (FE'ers) in attendance. Following on from that, Olshansky et al. [2020] conducted further interviews to learn about how people are converted to believing in FET. They found that most respondents only came to believe FET after watching YouTube videos. Various interviewees claimed to have been initially sceptical, but were eventually convinced after trying to disprove it themselves. For some, belief was a reinforcement of existing religious beliefs. Both scientific and religious arguments were key motivators. Many respondents already believed other conspiracy theories prior to FET, which is inline with previous work in conspiracy theories [Goertzel, 1994].

Landrum et al. [2021] performed a study of how susceptible people are to flat Earth videos on YouTube. They found that people with lower scientific intelligence and higher conspiracy mentality were more susceptible. Their findings also suggested that people found scientific arguments more compelling than religious ones. Mohammed [2019] carried out a basic content analysis of FE YouTube videos. Amongst videos on the

[3] https://www.bbc.co.uk/news/av/stories-49021903
[4] https://www.bbc.co.uk/news/blogs-trending-41399164

FE topic, there were more pro-FE than anti-FE videos, though the anti-FE videos were much more popular. Pro-FE videos were longer, and more likely to touch on religion and other conspiracy theories. Debunking videos were more likely to discuss science and maths, and reference established scientific works. Landrum and Olshansky [2020] performed an analysis of public perceptions of susceptibility to FET, and found that religious belief and level of education were predictors of susceptibility.

So far, the work in this area has focused on YouTube, which is understandable as it is the main way that people come into contact with FET. However, they have not looked at the dedicated FET communities, such as fora and subreddits. This may not be where non-believers are exposed and converted to FET, but it is where many members of the community gather to discuss ideas, and where round-earthers sometimes come to challenge them. Existing work has also not performed any linguistic analysis, or large scale content analysis of FE discussion. Another thing that has been overlooked is the mixed beliefs within the community. For example, many RE'ers participate in discussion on FE fora, not to mention trolls or insincere FE'ers. These are the problems we seek to discuss and address in this chapter.

6.3 Data Collection

In this research, the primary data source was a Flat Earth forum, `tfes.org`. To provide a brief history of the forum, in 2005 a forum called "`theflatearthsociety.org`", began as a place to discuss Flat Earth Theory. The forum "`tfes.org`" was started in 2013, and serves as our primary data source. This forum is very similar to the original, and in fact span off from its larger sibling, taking various members of the community with it. Both of these forums encourage debate of topics surrounding the Flat Earth. Members discuss the scientific basis of their claims, and often debate with "Round Earthers" on aspects of Flat Earth Theory. The forums welcome people with different beliefs on the Flat Earth.

Before creating our dataset, we considered the ethical implications of doing so. As a reference, we followed the guide to social media research provided by Townsend and Wallace [2016]. In previous works it has been considered reasonable to use data from public fora that do not require a login for access [Seale et al., 2010]. We think

it is reasonable to consider the forums in our study as public. In addition, our sharing and analysis of the data does not provide a risk of harm to any community members featured. For these reasons, we think that it is perfectly ethical to carry out analysis on this data. Even so, we did successfully receive ethics approval to conduct this research[5], conditional on the data being anonymised, and full quotations not being used in any published work.

To scrape the data from the forum, we used the python package Scrapy [Kouzis-Loukas, 2016]. The scraper used was a spider which followed every available link within the domain of the forum, until it ran out of links. For each page, it dumped the raw html to a file. By doing this, we were able to create a mostly complete snapshot of the forum at the time when we scraped. Because this data was collected retrospectively, any posts or users that have been deleted from the forum will not be present. Similarly, any information locked behind password protection was not scraped. The spiders gathered data in accordance with each site's *robot.txt*, with time delays built in so as not to harm the functionality of the sites by overwhelming them with requests.

Once a dump of HTML files had been created, the files were parsed to gather usable information. For this task, the Python package Beautiful Soup[6] was used. This allowed us to parse the HTML, pulling out useful information about each post, board, topic, and user. The retrieved information was sanity checked to ensure it was of a reasonable quality. When parsing the text of each post, quotes were replaced with a quote tag, which links the post to the quote's source. This was done so that quote text was not considered part of a post, while processing the text. This information was then recorded in an SQLite3 database[7]. Table 6.1 shows the data that was gathered from the forum.

6.3.1 Preprocessing and Tokenisation

Before text is analysed, it must be preprocessed and tokenised. For preprocessing, all quotes and URLs were first removed. These features should not be considered as part of the text, though it may be of use to record them separately, as they may give an impression of how posts are linked, and whether arguments are backed up with evidence, etc. The text was then normalised for unicode. The forums natively use

[5] We were not approved to use `theflatearthsociety.org`, for licensing reasons.
[6] `https://beautiful-soup-4.readthedocs.io/en/latest/`
[7] `https://sqlite.org`

Table 6.1: Table showing the data that was gathered from the FES Forum.

Post Information	
Property	Description
Post ID:	UID of the post.
Topic:	UID of the post's topic.
Topic Name:	Name of the post's topic.
Board:	UID of the post's board.
Board Name:	Name of the post's board.
User:	ID of poster.
Text:	Text of post, quotes replaced with tags.
Time:	Timestamp of post.
User Information	
Property	Description
ID:	UID of user.
Name:	Username at time of scraping. (removed for anonymisation)
Position:	Position on forum (e.g. 'moderator')
Custom Title:	An optional title, chosen by user.
Personal Text:	Text description to appear below user name.
Signature:	Text to follow a user's post.
Location:	Location of user.
Age:	Age of user.
Gender:	Gender of user.

Latin-1 encoding, however UTF-8 is the native encoding for python, and so we chose to convert the text to this new encoding. Finally, all repeated white-space was replaced with a single space. This was simply a tidying step, to help with tokenisation.

This preprocessing is quite minimal. We may consider going further, possibly removing punctuation, or numbers, etc. These are further steps we will experiment with, and they may each be useful for different tasks.

For tokenisation, we experimented with two different packages: Stanza [Qi et al., 2020a] and spaCy [Honnibal and Montani, 2017a]. Stanza is slightly more accurate, and provides slightly more features, while spaCy is significantly quicker. Both of the packages provide a pipeline that includes tokenisation, part-of-speech tagging[8], and named-entity recognition. Stanza also has the ability to find morphological features[9], as well as predicting the sentiment of sentences. Not all the pipeline needs to be used for most tasks. For example, PoS tags may only be needed when looking at language style, and Named Entities can provide insight into the people and places discussed, but are not

[8] https://universaldependencies.org/u/pos/
[9] https://universaldependencies.org/u/feat/index.html

needed in most tasks. The increased speed of spaCy means that it is often preferable to Stanza in the case of large datasets. Stanza was chosen for this work, due to the fact that it is more accurate than spaCy and provides some extra functionality [Qi et al., 2020a]. Computation time was not too much of a problem, as tokenisation could be run once, and the tokens used for all analyses.

6.3.2 Reddit Data Collection

Alongside the forum data, we also created a dataset of Flat Earth subreddits from popular social media website `reddit.com`. As of writing, Reddit is the 19th most popular site on the internet[10]. Reddit is a news aggregation site made up of sub-communities known as *subreddits*. Users make *submissions* to these subreddits, and then other users can *comment* on these submissions. The format is very similar to that of a typical online forum, making it a comparable source of data for our FES forum dataset.

Various subreddits exist dedicated to the discussion of Flat Earth belief. It is almost impossible to determine which of these, if any, are "genuine" so we opted to simply include a range of subreddits and accept that they may have differing views as to the legitimacy of Flat Earth theory. This is an issue that is inherent to researching conspiracy theories. However, as far as we are concerned, the language of conspiracy involves the way that they are discussed by non-believers as well as believers.

Eight subreddits were chosen, by searching reddit with the search term "flat earth". These were the eight subreddits that we found, with the highest number of contributions. While not comprehensive, the other FE subreddits we found were insignificantly small. We believe that eight provides sufficient variety. A meta analysis of these subreddits is performed in Section 6.5.

All eight subreddits were downloaded using the Pushshift API [Baumgartner et al., 2020]. This allowed us to gather every comment and submission on each of these subreddits. The data underwent the same preprocessing and tokenisation as described in Section 6.3.1, with the exception that quotes do not exist in the Reddit data. As with the FES forum, we created a database for each subreddit to make the data easily queryable.

One particularly interesting subreddit, which links to the previous section, is `r/`

[10] `https://alexa.com/topsites`

`flatearthsociety`: the sister subreddit to the FES forum. What is interesting about this subreddit is that it shut in 2017, blaming Round Earth believing trolls, who allegedly derailed the serious discussion. In this analysis it would be interesting to see if we can observe any aspect of this alleged behaviour in the community, or other Flat Earth communities.

One unknown aspect of these subreddits is how much overlap there is in their communities. Is there a general Reddit community interested in the Flat Earth who keep track of several subreddits, or does each subreddit have its own distinct community? This problem will not be investigated in this work, but would be an interesting avenue for future research, as long as rigorous ethical standards were followed. No effort was made in this thesis to link users between different platforms, or find users operating under several usernames, as this would breach standard ethical practice.

As well as comparing the FE forum to FE subreddits, we also looked at two non-FE subreddits as a comparison: `r/science` and `r/conspiracy`. These two were chosen as there were similar to Flat Earth communities in interesting ways: `r/science` relates because FE communities often discuss scientific experiments, and often engage in technical discussion, and `r/conspiracy` is similar because it is dedicated to conspiracy theories in general, one of which is the Flat Earth. It will be interesting to see which of these communities FE fora/subreddits are more similar to. Is there a language of conspiracy, and is pseudo scientific discussion similar to genuine scientific discussion? This data was collected in the same manner as the FE subreddits.

6.4 Meta Analysis

To better understand the forum, a thorough meta-analysis was performed to reveal interesting characteristics. Table 6.2 shows the basic statistics for `tfes.org`. The forum contains a substantial amount of text, spread across 126,200 posts. It is split into 13 boards, each representing a different section of the site. Examples of boards include, “Flat Earth Theory” and “Arts and Entertainment”. Within these boards are “topics”, which each contain a number of posts and encompass a single continuous thread of discussion. This is the structure typical of most fora.

Table 6.2: Table showing the basic meta-statistics of tfes.org.

Property	Value
# Posts	126,200
# Words	10,928,567
# Users	2319
# Boards	13
# Topics	5138

Table 6.3: Table showing the basic meta-statistics for members on tfes.org.

Property	Value
# Users	2318
Mean Posts / User	46.9
Median Posts/User	2
Mean Words/User	4048
Median Words/User	204.5

6.4.1 Distribution of Posts Across Boards

Figure 6.1 shows the number of posts on each board. 'Flat Earth Theory' has the most posts. This is unsurprising, as the forum primarily exists as a community in which to discuss the Flat Earth. What is more interesting is that neither of the two boards with the next most posts are specifically Flat Earth related. This is not to say that the Flat Earth never comes up in these boards, but it is nonetheless interesting that the forum is used as a more general social space, presumably one for those who believe the Earth is flat. Looking at words instead of posts (Figure 6.2) changes the order slightly, suggesting that the Arts and Entertainment board is made up of shorter posts. Overall, however, it gives a similar impression.

Based on this observation, boards have been split into three general groups: "Flat Earth", "Off-topic", and "Miscellaneous". These designations are the same as those in place on the forum itself. Flat Earth boards encompass discussion about different aspects of Flat Earth Theory, for example discussing theories or experiments. Off-topic discussion is more general, covering topics such as media, technology, and religion. Miscellaneous boards concern forum announcements and admin. When looking at the data with these groups in mind, Flat-Earth boards contain roughly 55% of the forum's posts, with off-topic and miscellaneous containing 40% and 4% respectively.

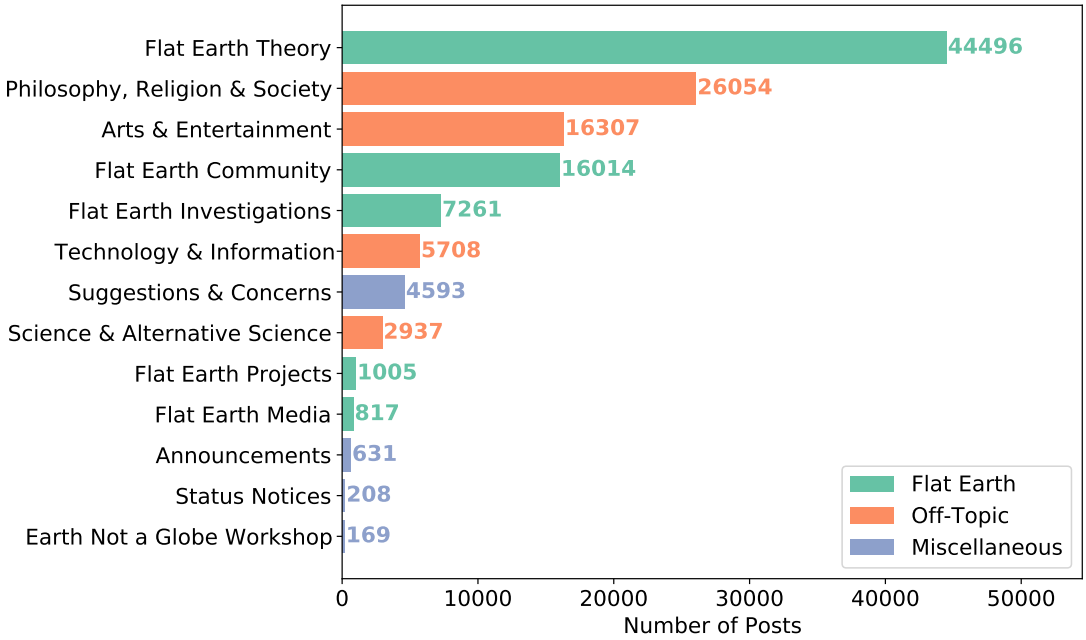


Figure 6.1: A bar plot showing the number of posts in each board in the tfes.org forum.

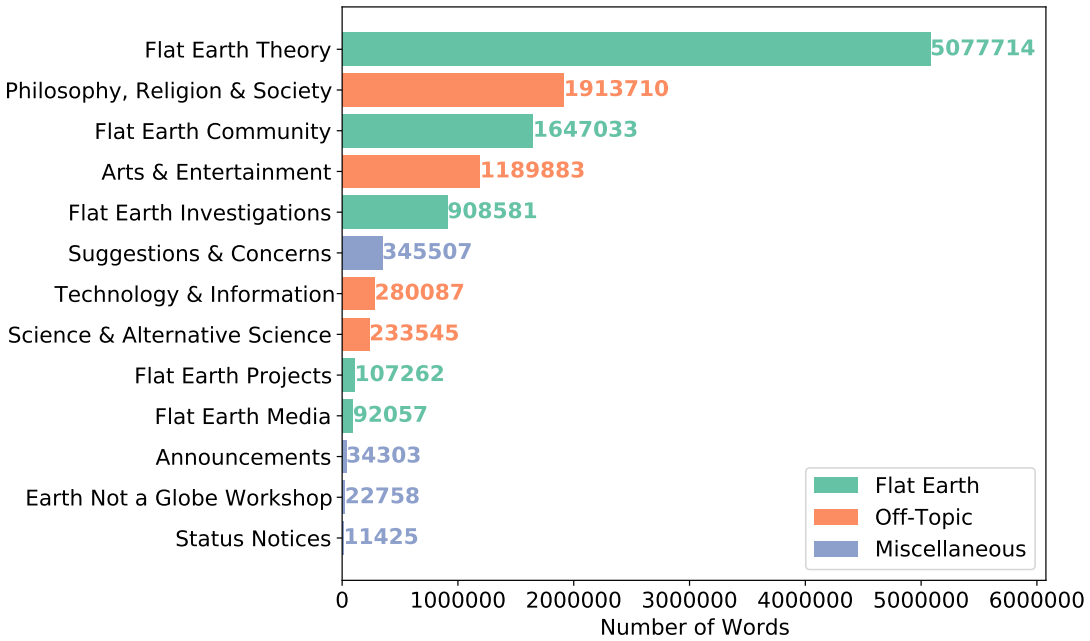


Figure 6.2: A bar plot showing the number of words in each board in the tfes.org forum.

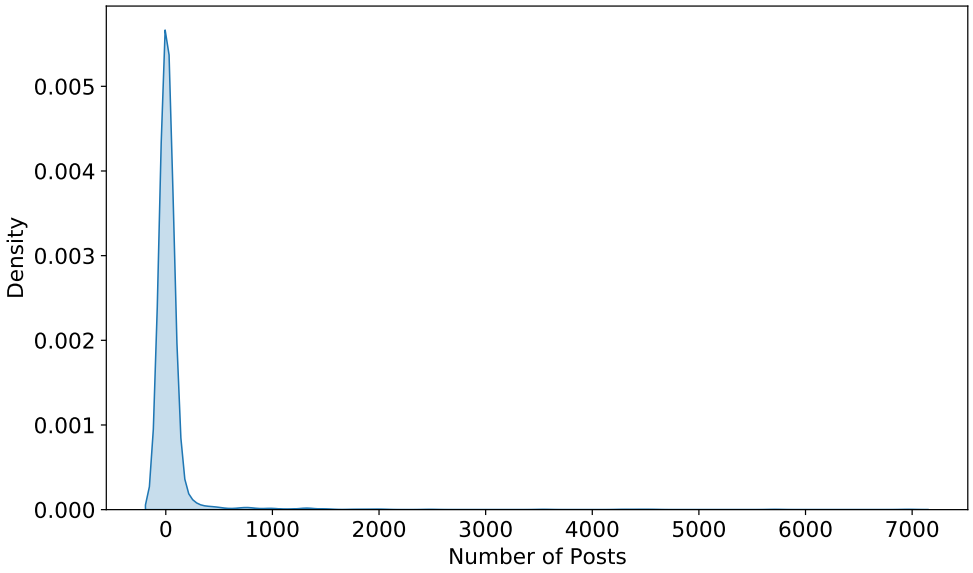


Figure 6.3: A density plot, showing the distribution of posts per user in the forum.

6.4.2 Distributions of Posts Across Users

Basic meta information about members of the forum is shown in Table 6.3. Figure 6.3 shows the distributions of posts per user on `tfes.org`. Immediately, it is clear that the vast majority of users contribute very few posts. In fact, out of 2,319 members, 964 posted once – that is roughly 42% of the forum's users. Only 439 members posted more than 10 times, and 142 more than 100. This points towards a very small active user-base, who make up the bulk of the forum. The meta-statistics also reflect this, with the median number of posts per user only being two, and the mean being skewed much higher.

Posts in the forum are concentrated amongst a small group of highly active users. The twenty users with the most posts are responsible for 48% of the posts on the forum. It is clearly important, therefore, to look at these users given how significant their contributions are. We will look more closely at these users in Section 7.4.

The forum has 12 users with roles. These are positions on the forum, e.g. moderator or administrator, which often involve responsibility for banning users, and keeping discussion within the rules of the forum. Role-holding members of the forum make up 23% of posts. As with the top twenty posters, which substantially overlaps with this group, looking at these members will clearly be important when trying to characterise

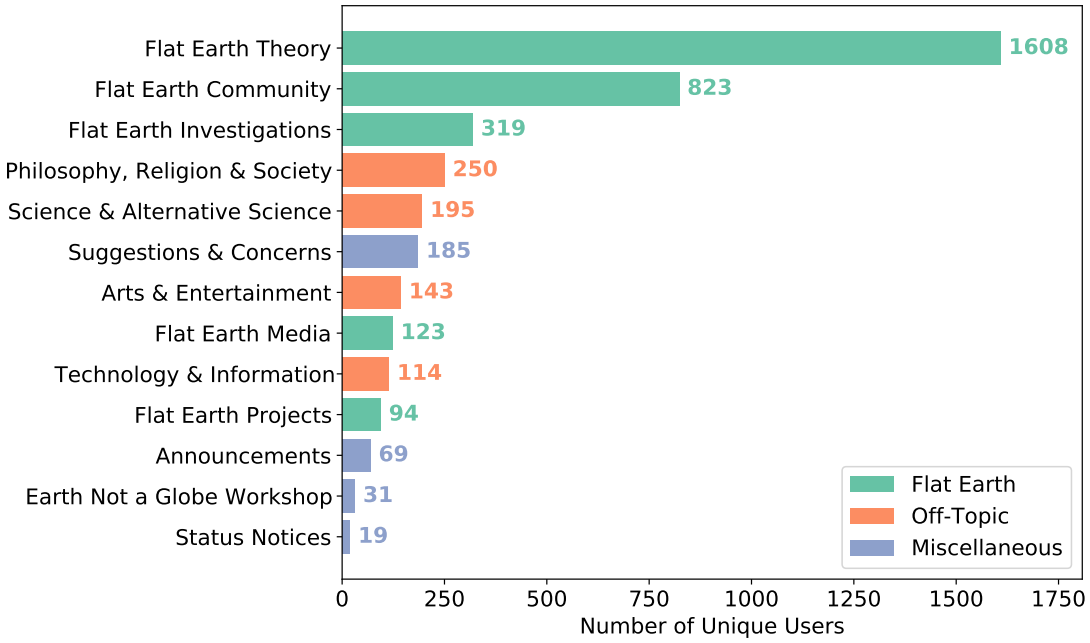


Figure 6.4: Bar plot, showing the number of unique users to post in each board.

the language of the forum.

There are a number of users with profiles that have been deleted. This means that their member uid's cannot be retrieved, and therefore they will not be included in our analysis. There is nothing that can be done about this problem. It is unfortunate, as it could result in many trolling users, as well as users who simply left, being excluded from the data. In an ideal world, we would stream posts live and keep everything ever posted. However, even if this were possible, it would be ethically dubious.

6.4.3 Distribution of Users Across Boards

We have already discussed the number of posts on each board, but it is also interesting to look at the spread of users across boards. Figure 6.4 shows the number of unique members who have posted in each board. Flat Earth related boards dominate here, with the three main Flat Earth boards having the most unique contributors. Off-topic boards appear to have fewer members. This is interesting, given that Figure 6.1 showed that "Philosophy, Religion & Society" and "Arts & Entertainment" were the boards with the second and third most posts. Dedicated members may be the only ones who use these parts of the forum. It would make sense that new users to a Flat Earth forum would go there to participate in Flat Earth discussion.

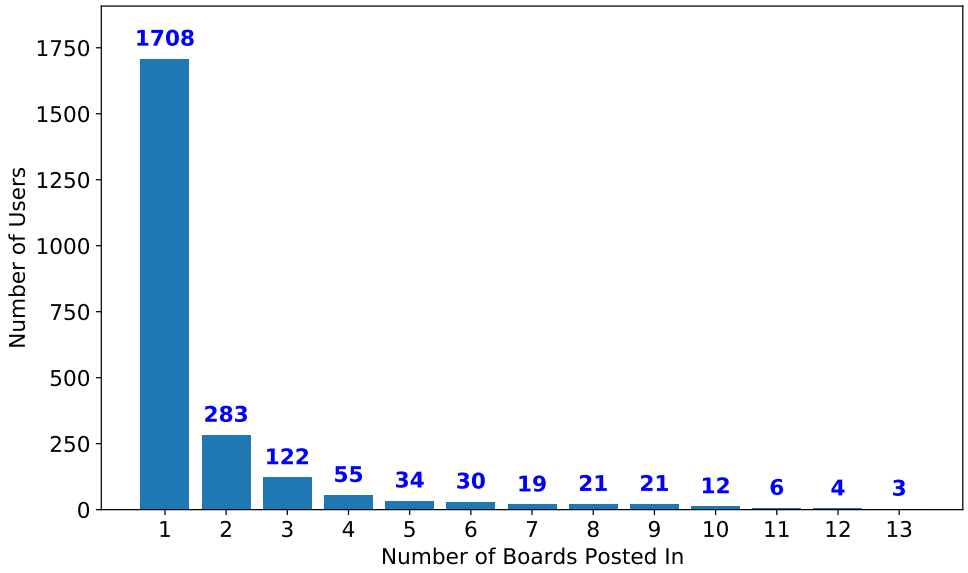


Figure 6.5: Bar plot, showing the number of users that have posted in each different number of boards.

This characteristic of the forum can be further investigated by looking at the users who post in the off-topic sections. 391 members (∼17% of users) have posted in off-topic boards. These individuals wrote ∼94,000 posts – roughly 87% of the forum's posts[11], providing further evidence to our suspicion that more regular users of the forum are more likely to engage in off-topic discussion.

Another useful way to look at the spread of users across boards, is to plot the number of boards each user has posted in, this is shown in Figure 6.5. The majority of users have only posted in a single board, and 2113 users have posted in no more than three. As is becoming a common theme, the users who posted in more than three boards were responsible for 87% of posts. This is the same proportion of posts by users who posted in off-topic.

6.4.4 Distribution of Posts Over Time

An important element to look at to better understand the forum is the number of posts over time. For example, it would be interesting to know if the forum undergoes ebbs and flows of activity, or whether it is quite consistent.

Figure 6.6 shows a rolling plot of the number of posts, with a window size of 90

[11]Excluding posts by deleted members. Including these posts, it would be ∼75%.

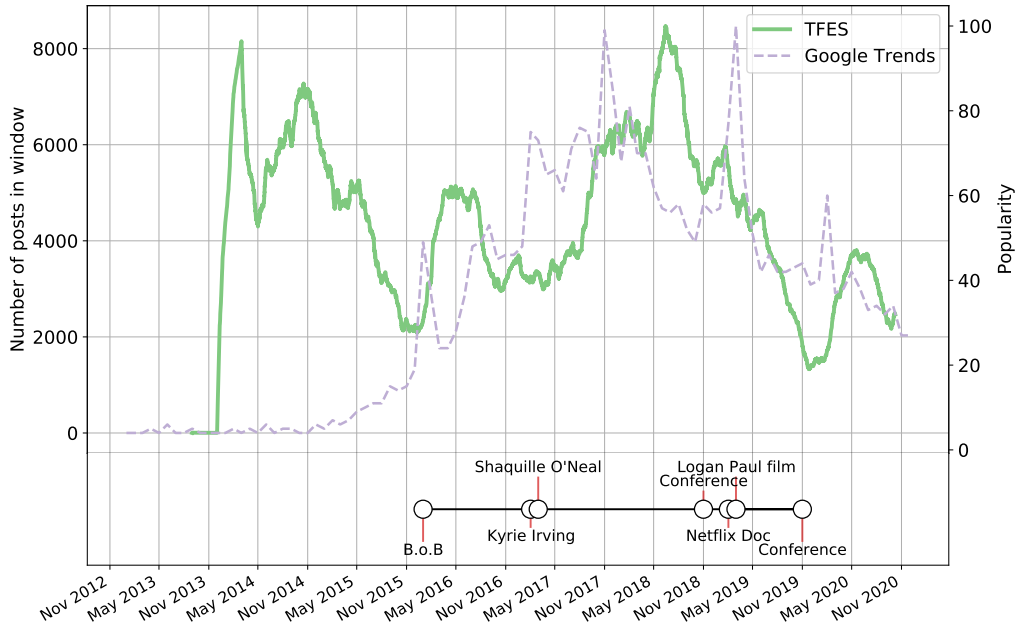


Figure 6.6: Graph of number of posts with a 90 day rolling window. Dashed line shows interest in the Flat Earth topic according to Google Trends. Some key FE events are marked along the bottom of the graph.

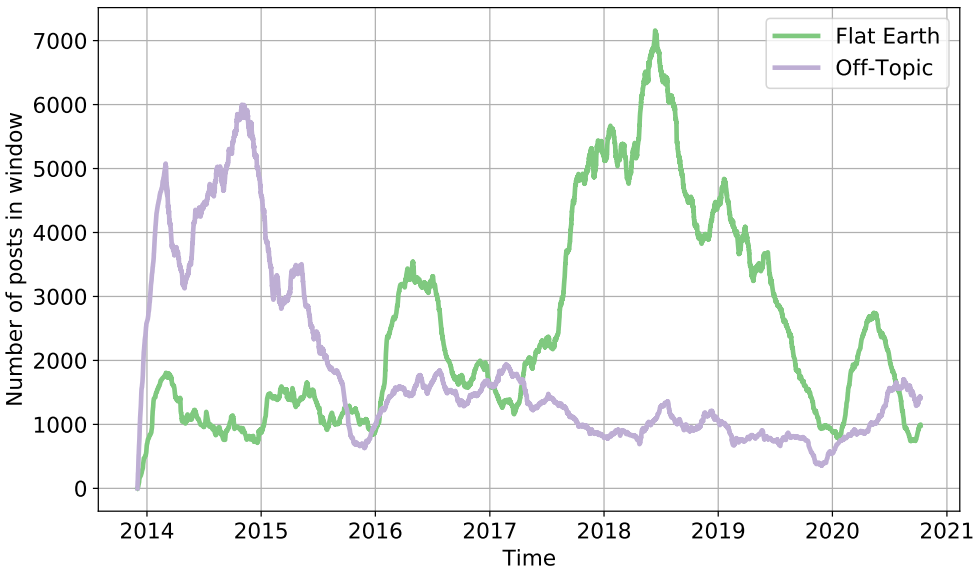


Figure 6.7: Graph of number of posts with a 90 day rolling window, showing the number of posts for Flat Earth and Off-Topic sections of the forum.

days. From this one can see that there are clear peaks in posting on the forum; an initial peak, followed by a small peak in 2016, and then a large one in 2018. It is possible that these correspond to times when Flat Earth went viral in the wider web.

Figure 6.6 also shows the popularity of the "Flat Earth" topic according to Google Trends[12] alongside the posts over time. This will give an impression of how many people searched Google with FE related queries over time, which gives an impression of wider popularity. As we can see, the second spike seems to correspond fairly well to the peak of popularity on Google. In fact, it seems to lag slightly behind the Google trend line. This could suggest that this peak was motivated by people finding out about FET and joining the community. The first peak seems to be less driven by wider popularity, which could mean that the core community can be identified by looking at the period between 2013 and 2016.

We also plotted several key FE events along the bottom of the graph. These events all take place during the period of popularity. Very early in its surge, B.o.B, an American rapper, voiced his belief in the Flat Earth. Following this event there was a spike in contributions. This could suggest how the popularity of the community is driven by external events. Several of the later events appear to take place following peaks in popularity. This may, therefore, suggest that these events were reactive to FE's popularity rather than the cause of it.

Figure 6.7 shows the number of posts over time from Flat Earth and off-topic sections of the forum. It highlights how the FE sections have massively overtaken the off-topic parts. When the forum started, it seems to have been used as more of a general social space, possibly for FE believers. As the FE topic became more popular in the wider web, however, the overwhelming majority of posts were to the FE sections. This gives us further indication that the users active before 2016 may be considered the core community.

6.4.5 How Long Do Users Stay?

An interesting question to investigate is how long members stay on the forum. This is difficult to know for sure from this data, as we have no record of all the users that do not post. However, as we are looking at the language of the forum, and such users do

[12] `https://trends.google.com`

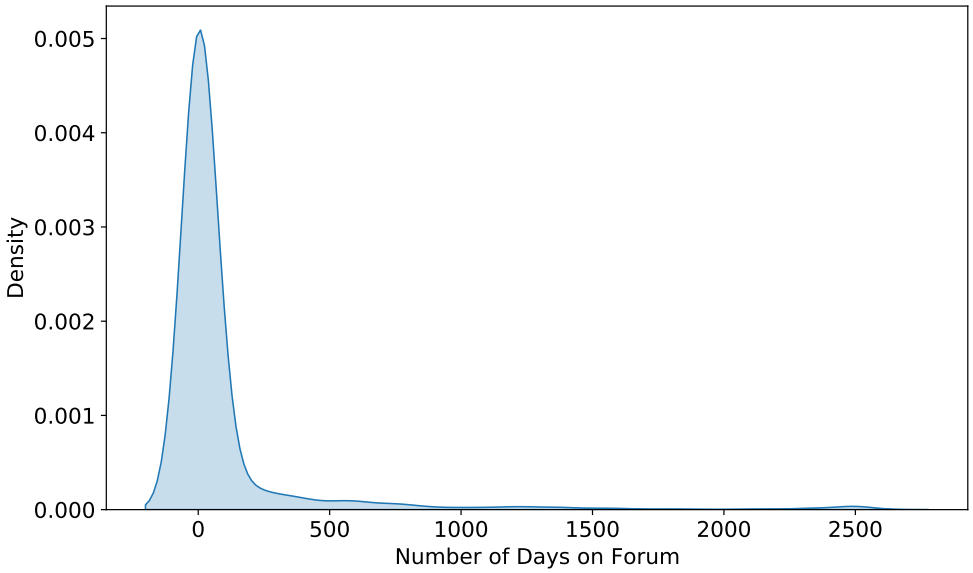


Figure 6.8: Density plot showing the length of time between the first and last posts for users on the forum.

not contribute to this language, a member's lifetime is considered as the number of days between their first and last post.

Figure 6.8 shows the distribution of user lifetimes. The vast majority of users stay on the forum for a very short span of time. 60% of members were only active for a single day. A much smaller number of users have been active for a long period of time. 24 members were active for more than 5 years. These members contributed 38% of the posts on the forum.

6.4.6 New Users Over Time

Figure 6.9 shows a plot of the number of new users over time on the forum. A user is counted as new on the date of their first post. The graph shows a similar peak in the forum's popularity, although slightly earlier in 2018. There are several brief spikes in new members throughout the period, suggesting that new users come in waves, possibly when FET is trending. In the graph, we included some key events relating to FET. The spikes do not all correspond neatly to events, namely the biggest spike of all, though it follows Flat Earth being in the news thanks to celebrity endorsements from Shaquille O'Neal and Kyrie Irving.

Another observation from this graph is that there are comparatively very few new

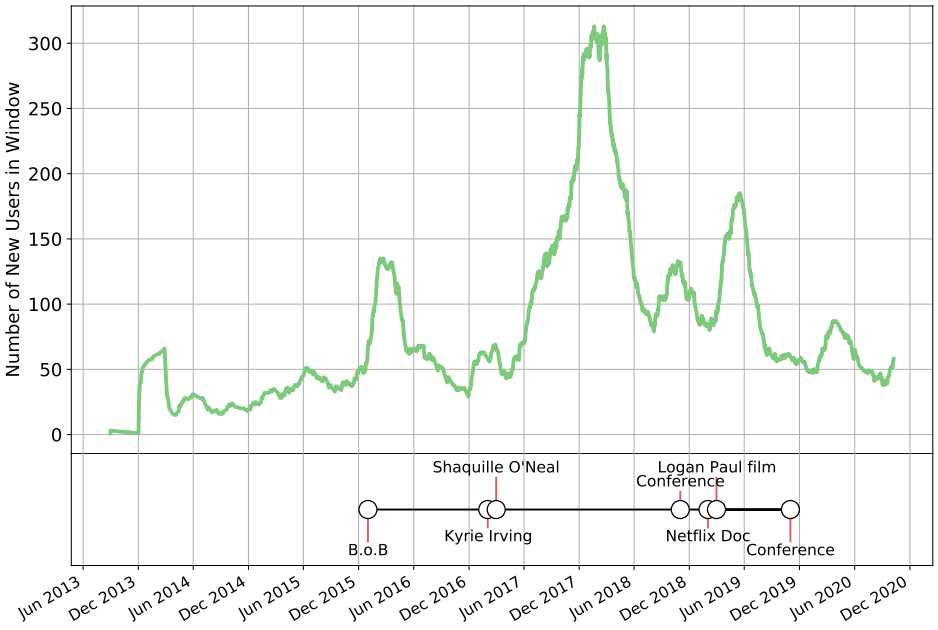


Figure 6.9: Plot of the number of new users over time, according to date of first post.

users prior to the first spike in 2016 (after B.o.B announced his FE belief). This could suggest that looking at members who were active before this point may allow us to identify the core FE community, rather than members who joined following the topic's mainstream popularity.

6.4.7 Main Insights from Meta-Analysis

In this section a meta-analysis has been conducted, highlighting various interesting aspects of posting and user behaviour on the forum. We found that a large proportion of the posts on this forum are contributed by one-time users, and a small group of dedicated members do the bulk of the posting. This behaviour is typical of online communities [van Mierlo, 2014, Glenski et al., 2017], but by looking at the differences in language between these two groups of people, we may be able to answer interesting questions about conspiracy communities. For example, it would be valuable to see whether these ephemeral visitors are trolls, people coming to mock the theory, or genuinely curious individuals. Along similar lines, do the dedicated users all voice support of Flat Earth Theory, or do sceptics also hang around?

Another interesting observation has been that a relatively active minority of users

Table 6.4: Table showing the basic meta-statistics for the Flat Earth subreddits we are looking at.

Subreddit	Num Comments	Num Submissions	Unique Users	Start Date
DebateFlatEarth	11,224	649	941	12/10/2016
flatearth	498,262	44,622	32,645	19/01/2013
FlatEarthIsReal	2,599	475	532	18/07/2017
flatearthsociety	17,989	1,128	1,835	09/06/2012
Flat_Earth	27,005	2,094	3,873	08/05/2013
Globeskeptic	15,171	1,692	1,833	26/02/2020
notaglobe	27,883	2,175	5,325	25/08/2018
theworldisflat	31,456	1,659	4,827	17/06/2015

post in the off-topic sections of the forum. This might be a helpful way of identifying groups of dedicated members. By looking at the spread of boards that users post in, one might be able to separate those who are core members of the community from visitors and trolls.

Already, just from the meta information, several ways for identifying potential groups of users have been identified. By looking at the language of these groups, with methods such as keywords analysis, the commonalities within the groupings may become apparent.

6.5 Subreddit Meta-Analysis

In Section 6.4, we carried out a meta-analysis of the Flat Earth Society Forum. In this section, we will perform a meta-analysis on a set of flat Earth, as well as two non-flat-Earth, subreddits. Because Reddit is a much more popular, mainstream website (certainly compared to the `tfes.org`), we hope that this will provide a slightly different context for flat earth discussion.

6.5.1 Flat Earth Subreddits

Following on from our meta-analysis of the FES forum, we will now perform a similar analysis with the FE subreddits we described in Section 6.3.2.

Basic Statistics

Table 6.4 shows some basic meta-statistics for the subreddits we are looking at. There is a large variation in the size of the subreddits. The smallest subreddit contains only 2,599 comments, and the largest 498,262. The second largest has around 30,000, meaning that `r/flatearth` is by far the biggest.

The oldest FE subreddit has been running since 2012, but subreddits have been consistently created over time. It is possible some of these communities span out of existing subreddits, as often happens with reddit communities[13] [Hessel et al., 2016].

On some of the subreddits, users have flair text. This is text that appears beside their username. Some subreddits have useful flair text, such as `r/theflatearthsociety`, where users have flair text such as: "Flat Earth", "Round Earth", and "Undecided". In examples like this, these flair texts could be used as labels of belief (or at least purported belief). On most of the subreddits, however, users seems to use flair text for more humorous purposes, so it will not always be useful.

Contributions Per User

Looking at the comments per user (Table 6.5), it is noteworthy that the median number of comments for a user is 2 for most subreddits. For submissions per user (Table 6.6), the median is 1, although it is worth noting that this only includes users who do make a submission. This suggests that most users do not make many posts, as was the case on the forum.

The disparities between the number of posts are enormous. A single user on `r/flatearth` posted 9,000 times.

Some subreddits have more activity than others. The mean posts per user is much higher for the largest subreddit. We would need to divide this by time, however to know whether this is down to more active users, or the subreddit having existed for longer.

User Lifetimes

Table 6.7 shows the average lifetimes of users on the FE subreddits. The majority of users only spend a single day on the subreddits. All of the communities seem to have a

[13] And online communities in general, as was the case with the two Flat Earth fora.

Table 6.5: Table showing the number of comments per user on each subreddit. Shows the mean, 25th, 50th, and 75th percentiles, and maximum.

		Percentile			
Subreddit	**Mean**	**25%**	**50%**	**75%**	**Max**
DebateFlatEarth	11.70	1	2	6	699
flatearth	14.36	1	2	4	9,697
FlatEarthIsReal	4.77	1	2	4	200
flatearthsociety	8.22	1	2	5	1,036
Flat_Earth	6.70	1	2	4	846
Globeskeptic	7.66	1	2	5	968
notaglobe	4.78	1	1	3	1,499
theworldisflat	4.33	1	1	2	1,630

Table 6.6: Table showing the number of submissions per user on each subreddit. Shows the mean, 25th, 50th, and 75th percentiles, and maximum.

		Percentile			
Subreddit	**Mean**	**25%**	**50%**	**75%**	**Max**
DebateFlatEarth	1.76	1	1	1	41
flatearth	2.27	1	1	1	1,411
FlatEarthIsReal	1.61	1	1	1	20
flatearthsociety	1.43	1	1	1	41
Flat_Earth	1.28	1	1	1	28
Globeskeptic	2.37	1	1	1	223
notaglobe	11.02	1	1	3	838
theworldisflat	7.00	1	1	2	253

Table 6.7: Table showing the lifetimes of users in days, on each subreddit. Shows the mean, 25th, 50th, and 75th percentiles, and maximum.

		Percentile			
Subreddit	**Mean**	**25%**	**50%**	**75%**	**Max**
DebateFlatEarth	99.57	1	1	39	1,522
flatearth	68.74	1	1	16	1,942
FlatEarthIsReal	29.86	1	1	4	1,261
flatearthsociety	23.55	1	1	2	1,017
Flat_Earth	57.84	1	1	9	1,762
Globeskeptic	12.80	1	1	4	310
notaglobe	26.96	1	1	1	849
theworldisflat	28.39	1	1	1	1,906

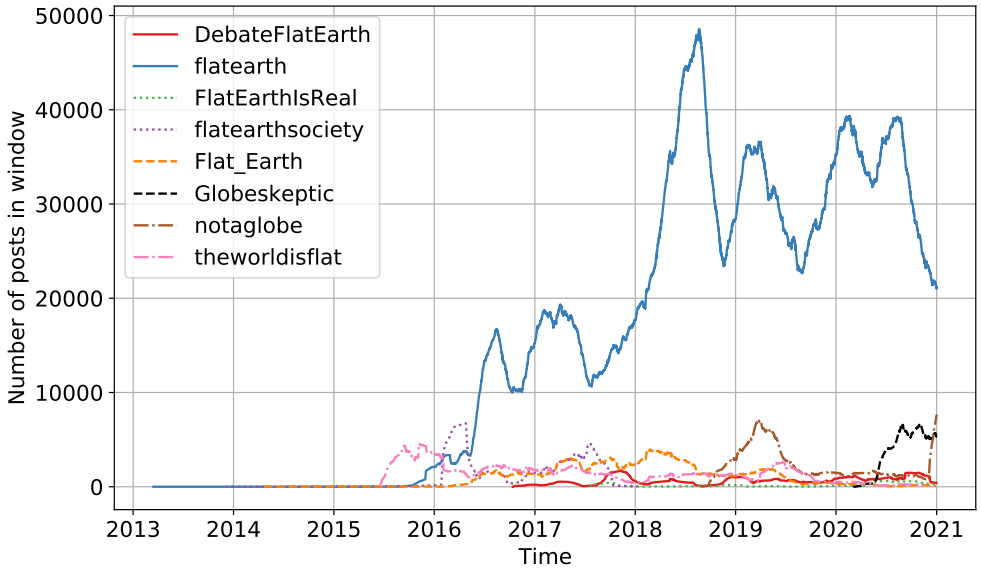


Figure 6.10: Plot of the number of comments over time on the Flat Earth subreddits.

small number of users active for a long time. These could be core community members and moderators, etc. This is similar to what we found looking at the FES Forum.

One comparatively small subreddit, `r/DebateFlatEarth`, had the longest average user lifetime. This could suggest that this subreddit has a more dedicated community than the others.

Contributions Over Time

Figure 6.10 plots the number of comments made over time on the subreddits in question. We also plotted the same graph, excluding the largest forum to make the smaller ones clear. This can be seen in Figure 6.11. This can give us an impression of how posting behaviour changed in these communities. Here, we will discuss mainly comments over time, but submissions over time tell us a very similar story.

The subreddits we are looking at began from 2013 onwards, but it is not until late 2015/early 2016 that comments in these communities really got going. Three of the communities began before this increase in popularity: `r/flatearth`, `r/flatearthsociety`, and `r/Flat_Earth`. It is possible that these communities were formed before the Flat Earth became mainstream[14]. Whether or not this means that they are more likely to be "genuine" is not clear.

[14]As shown earlier in Figure 6.6, interest in the Flat Earth on Google only picked up in 2015.

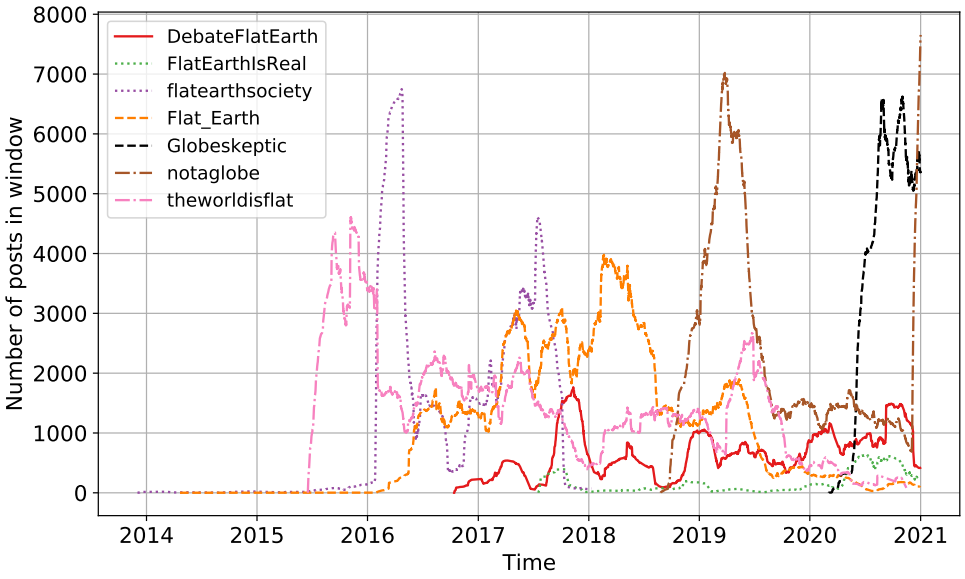


Figure 6.11: Plot of the number of comments over time on the Flat Earth subreddits. `r/flatearth` has been removed due to it being so much larger.

Various subreddits enjoy an initial peak following their creation, and then a large drop off. This could be a feature of these types of online communities, that they become briefly popular, but that attention does not last. It would also be interesting if the large initial peaks were caused by users migrating from a different community.

There are large peaks in activity, and activity in general is fluctuous. The peaks may line up to points in time when Flat Earth became popular online. What may suggest otherwise is that the peaks do not line up particularly well between subreddits. This may point to these peaks applying to particular communities rather than general interest in the conspiracy.

To look specifically at `r/flatearthsociety`, it is interesting that it has two large peaks in activity, one in early 2016, and another in early 2017. Following this second peak, the subreddit falls rapidly and then ceases to exist, in mid 2017. This seems to play into what was said on the forum: that an influx of undesirable users had flooded in and forced them to retreat to their forum. Interestingly, two other subreddits (`r/Flat_Earth` and `r/DebateFlatEarth`) have increases in comments around the same time as `r/flatearthsociety`'s demise. These could have been fuelled by users leaving the dying subreddit. A small subreddit, `r/FlatEarthIsReal` starts around the same time as `r/flatearthsociety` ceased to be. This community may

Table 6.8: Table showing the number of removed comments and submissions for the Flat Earth subreddits we are looking at.

Subreddit	Removed Comments	Removed Submissions
DebateFlatEarth	78 (0.69%)	18 (2.77%)
flatearth	15,986 (3.21%)	3,252 (7.29%)
FlatEarthIsReal	14 (0.54%)	6 (1.26%)
flatearthsociety	445 (2.47%)	108 (9.57%)
Flat_Earth	338 (1.25%)	122 (5.83%)
Globeskeptic	822 (5.42%)	319 (18.85%)
notaglobe	2,567 (9.21%)	12 (0.55%)
theworldisflat	9,448 (30.04%)	65 (3.92%)

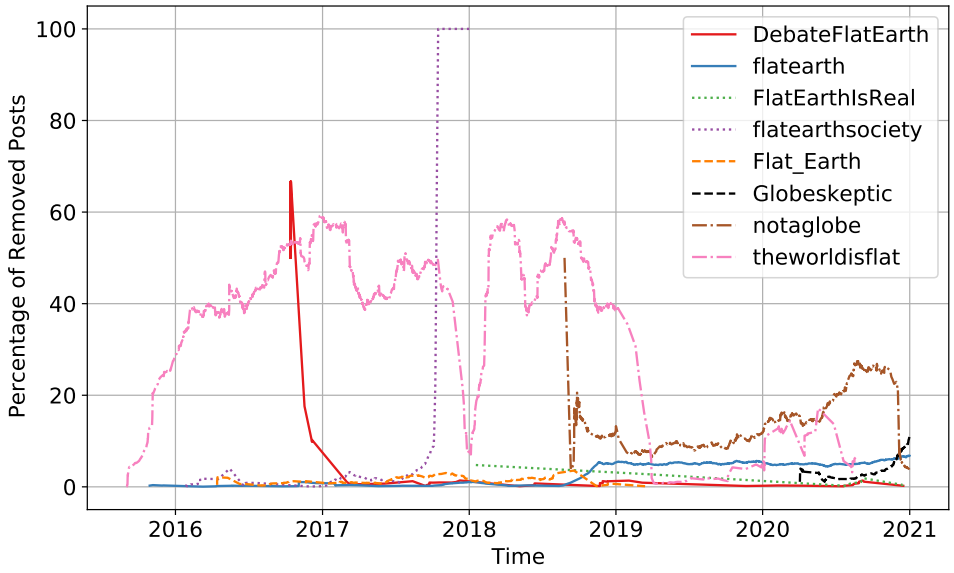


Figure 6.12: Plot of the proportion of all comments that were removed over time on the Flat Earth subreddits.

have been started by ex-members.

The death of `r/flatearthsociety` also roughly corresponds to an increase in users on the FES forum throughout 2017, as shown in Figure 6.6. This increase could either be driven by the same factors that allegedly affected the subreddit, that of visiting Round Earthers, or it could suggest that users from the declining subreddit moved back to the forum. It is difficult to assess which of these may be the case without attempting to link users between the two communities, which we will not attempt for ethical reasons.

Removed Contributions

Reddit posts sometimes have a flag to indicate that they were either deleted or removed. "Deleted" tends to imply that a user has deleted the post themselves, and "removed" suggests that it was taken down by a moderator. By looking at the proportion of contributions which were removed over time, we can get an impression of various things. For example, it may suggest the moderation getting stricter, or a surge of new users breaking a sub's rules. Table 6.8 shows the number of removed comments and submissions on each of the subreddits we are looking at. Figure 6.12 shows the proportion of comments or submissions where the contents of the message has been removed over time.

One notable subreddit here is `r/theworldisflat`, which at one point was deleting almost 60% of its comments. The proportion of removed posts is sustained at a high level for quite some time, which may suggest that this is down to forum rules/strict moderation rather than a particular surge in unscrupulous activity.

Looking at `r/flatearthsociety`, one can see that just before it ends, there is a huge surge in deleted comments, culminating in 100% of comments being deleted when the community ends its life. In terms of removed submissions, there are two peaks of deletion, one in early 2016, and one running up to the sub's death. These line up with the subreddit's popularity, and suggest that there is some truth in the administrator's allegation that when large groups of users turned up, many posts broke the community guidelines.

6.5.2 Comparison Subreddits

In this section we will perform a meta-analysis of the two non-FE communities we introduced in Section 6.3.2.

Basic Statistics

Table 6.9 shows some basic meta-statistics for the two non-FE subreddits. There are more total contributions on `r/conspiracy`, but fewer active users. This points to a smaller but more active community.

As shown in Table 6.10 `r/conspiracy` has fewer removed posts. 18% of

Table 6.9: Table showing the basic statistics for the comparison subreddits.

Subreddit	Number of Comments	Number of Submissions	# Unique Commenters	Start Date
conspiracy	18,712,904	1,072,166	620,277	29/01/2008
science	12,230,919	808,821	1,393,456	18/10/2006

Table 6.10: Table showing the number and percentages of removed contributions for the comparison subreddits.

Subreddit	Removed Comments	Removed Submissions
conspiracy	361,739 (1.93%)	32,362 (3.02%)
science	2,193,602 (17.93%)	4,454 (0.55%)

comments on `r/science` were removed, compared to 2% on `r/conspiracy`. This could point towards lighter moderation on `r/conspiracy`, but could also be for another reason as it is not always clear why a post is removed.

Contributions Per User

Table 6.11 shows the average contributions for user on each comparison subreddit. Both subreddits have a median of 2 comments per user, which indicates that, as with all the other subreddits, most users only post very little. This suggests that this is not a feature of Flat-Earth debate, but rather a feature of online discussion on fora/Reddit in general. Despite the median being similar, `r/conspiracy` has a higher mean number of comments per user. This may mean that this subreddit has more active users. For submissions, the distribution was much the same.

User Lifetimes

Table 6.13 shows the average lifetimes for the comparison subreddits. The median user lifetime for both subreddits is one day, but seems to increase quite quickly afterwards. The 75th percentile is 390 for `r/science` and 230 for `r/conspiracy`. This still means that users who stick around for a significant amount of the subreddits' lifetimes are relatively few and far between. The mean user lifetime for `r/science` is higher. This may suggest a more stable, longer lasting community. The maximum lifetime of users in both is quite high: 13 years for `r/conspiracy`, and 14 years for `r/science`. For both, this is more or less the same as the lifetime of the entire subreddit, suggesting that in both communities, there are users who are active the whole time, and

Table 6.11: Table showing the number of comments per user for the comparison subreddits.

		Percentile			
Subreddit	**Mean**	**25%**	**50%**	**75%**	**Max**
conspiracy	27.45	1	2	8	65,272
science	6.09	1	2	4	22,203

Table 6.12: Table showing the number of submissions per user for the comparison subreddits.

		Percentile			
Subreddit	**Mean**	**25%**	**50%**	**75%**	**Max**
conspiracy	5.23	1	1	2	10,600
science	3.00	1	1	2	5,519

Table 6.13: Table showing the user lifetimes for the comparison subreddits.

		Percentile			
Subreddit	**Mean**	**25%**	**50%**	**75%**	**Max**
conspiracy	258.32	0	0	230	4,659
science	372.48	0	0	390	5,203

that the communities still contain some of their original members.

Contributions Over Time

Section 6.5.2 shows the contributions over time for both these subreddits, between 2013 and 2020, to replicate the time range of the FE communities. `r/science` is steady, increasing the entire time. `r/conspiracy` begins later than `r/science`, and initially has fewer comments in each window. By 2014, it had caught up with the number of comments per window of `r/science`. From then it follows a similar trajectory, which could possibly be down to general growth in Reddit usage, until a big surge in mid 2016, peaking in 2017. This lines up with various events that brought many conspiracies into the mainstream. For example, the 2016 presidential election brought with it conspiracies such as QAnon and Pizzagate. The number of comments pick up again in 2019 (possibly Donald Trump's impeachment?) but then increases massively in 2020. At the point where the data ends, it still seems to be increasing. It is possible that Covid-19 and the 2020 US Presidential Elections have prompted a resurgence in Conspiracy discussion.

It is interesting how large `r/conspiracy` grew to be compared to `r/science`.

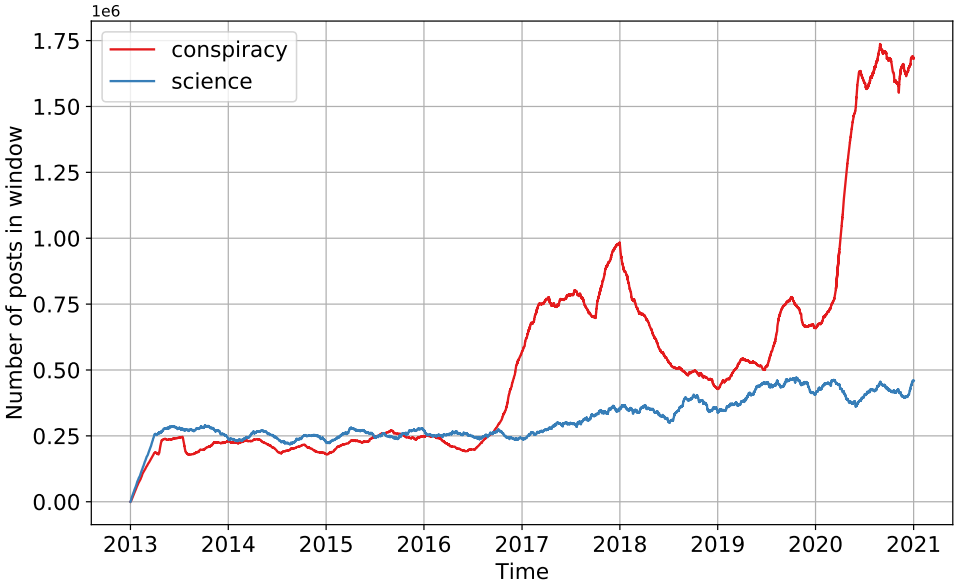


Figure 6.13: Plot of the number of comments over time on the comparison subreddits.

Given the popularity of some conspiracy theories [Li et al., 2020a, Papasavva et al., 2021], and the disproportionate influence of certain fringe communities on the web at large [Zannettou et al., 2017], this may not come as a surprise. But the meteoric rise of `r/conspiracy` around 2017, to become considerably more popular than a traditionally mainstream topic such as science, is an interesting change, nonetheless. Perhaps it is indicative of trends of internet discussion, or it could reflect a wider change in society towards engaging more with conspiracy theories.

Looking at submissions over time gives a very similar impression. Though, unlike with comments, submissions over time on `r/science` appear to decrease from 2014. It is not clear why submissions would decrease while comments increase.

Removed Contributions

Figure 6.14 shows the proportion of removed contributions on these two subreddits over time. `r/conspiracy` is fairly steady, though the proportion of removed comments appears to increase gradually over time. There are some increases, for example in 2018, just after the big peak of posting. A dip in removed comments at the same time as the peak of commenting could suggest that moderators could not keep up with the increasing number of comments. `r/science` has a very high proportion of removed posts from around 2016. This could possibly indicate more rule breakers, stricter

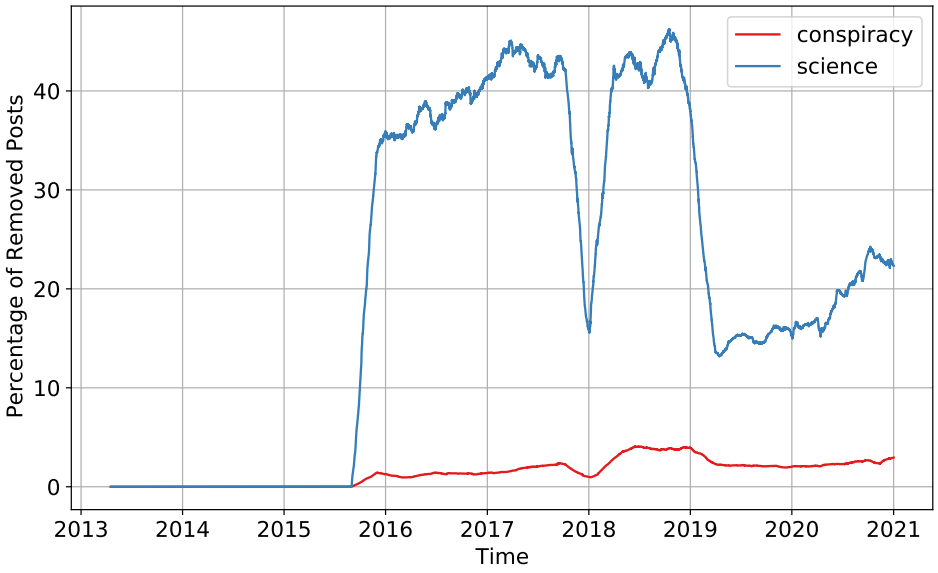


Figure 6.14: Plot of the proportion of removed comments over time on the comparison subreddits.

moderation, or it could just be a coincidence. Looking at deleted submissions, there is a large gap. This is quite possibly a data issue rather than a feature of the community.

6.5.3 Problems with Reddit data

As we have noted throughout this section, it can be difficult to know the completeness of Reddit data. There are known problems with the Pushshift dataset [Gaffney and Matias, 2018], which mean that one cannot assume completeness. Gaps have been present in our data. For example there is a strange dearth of comments between 2009 and 2011. Though this is outside the time range we are looking at it is still concerning. Missing data makes it very difficult to know for sure if a change is genuine or down to gaps in the dataset. For this reason, when performing analyses using the reddit data in Chapter 7, we will produce a sample with a consistent number of posts over time. This will only be done for the two comparison subreddits, as some of the FE communities do not have enough data to spare. The samples were made up of 100,000 comments and 10,000 submissions from each full 365 day window between January 2012 and January 2021, for each of the two subreddits.

6.6 What Characterises Flat Earth Debate?

Understanding the language used in Flat Earth Discussion is crucial to understanding the community. Topics discussed, usage of technical language, and the authorial style of community members can all tell us something interesting about the discussion of conspiracy theories online. In this section we will look at the linguistic features of flat-earth discussion compared to normal online conversation. This will be done using a range of feature sets, each designed to look at a different aspect of the community's language. Most of the features we discuss have been introduced previously in Chapters 2 to 4. They are predominantly content-based features, and will provide a general overview of differences in language usage.

To find the features that characterise Flat Earth debate, we compared the text in the FE sections of the forum to the off-topic sections. This will allow us to identify features which vary between these sections. Because of the need for self-labelled topics, this analysis will only be applied to the FES forum.

6.6.1 Keyness Analysis

Much of the following analysis, will involve finding "key" features, which differ significantly in one corpus compared to another. In this case we are looking at differences between Flat Earth Discussion and 'normal' off-topic discussion. In the example of the FES forum, Flat Earth and Off-topic were defined based on their board, as described in Section 6.4.1. By finding features that are over-used in Flat Earth discussion compared to Off-Topic, we can hope to identify the key features of Flat Earth discussion.

To identify key features, two techniques are employed which will be familiar from Section 2.2.1. The first is log-likelihood [Rayson and Garside, 2000], a significance statistic, that can be used to find features that are significantly overused in one corpus compared to another. Log-Ratio[15] is the other technique. This is an effect-size statistic that says how much more a term is used in one corpus than another. For each term, both values are calculated. All features that have a log-likelihood of > 3.84 are significantly different to a significance level of 0.05. Features were then ranked according to their

[15] http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/

log-ratio.

Sometimes one might put some extra limits on features before considering them eligible. For example, in the case of words, one might allow only terms that appeared more than 100 times in either corpus. This may help avoid rare words coming up as key, which could make observations more generalisable.

So far we have mentioned words. However, this analysis could just as easily be performed on almost any feature [Rayson, 2008], as it is simply a significance test followed up with an effect size measure. We can, for example, find key PoS tags, or key word trigrams, as we will demonstrate over the next section.

6.6.2 Word N-Grams

Key Unigrams

In this analysis, the aim is to identify words that occur significantly more in one part of the corpus than another. Specifically in this case, it will mean identifying words that are used significantly more on Flat Earth boards than on off-topic boards. This will provide an impression of the words that are used in Flat Earth discussion, that would not be used in normal conversation.

We began by looking at the words with the highest log-ratio scores, meaning they were overused on the FE section. Technical words relating to FET were, unsurprisingly, key. Words such as “longitude”, “circumference”, “refraction”. While it may not be particularly surprising that these words would be overused in flat earth discussion, it is interesting to get an impression of the key points of Flat Earth Theory, and what topics debate centres around. These words in particular give an impression that the discussion is not as shallow as simply stating that the Earth is flat, but rather that members will discuss how this theory relates to other phenomena.

These words being key also suggests that the Flat Earth is not discussed nearly as much in the off-topic section. This might make it more likely that only dedicated members of the community would post there. It seems unlikely that a troll would come along to a Flat Earth forum, only to provide their opinions on the latest Star Wars film.

Most of the words mentioned so far barely appeared in the off-topic sections. However, we are also interested in more general words that are overused in FE discussion. To account for this, we looked at key words with a minimum frequency

of 500 in both sections of the corpus. More common words such as "evidence", and "claim" are now highlighted as key. These still point to the idea that FE discussion involves a scientific style of discussion.

Keywords of the off-topic section focused around more general topics. Notable topics covered were: video games ("Morrowind", "Skyrim"), politics ("Biden", "Democrats"), and music ("vocals", "lyrics"). This seems like a typical range of topics that might be discussed in any range of online communities, and suggests that the forum does function as a general social space as well as a place to debate and share theories. Many more pronouns ("her", "she", "him") appeared as key in the off-topic section, which could indicate a more casual style of discussion.

Longer Word N-Grams

Looking at keywords can help identify words that are overused in Flat Earth discussion. However, the words highlighted are robbed of context, so it is hard to tell if words are significant by themselves or if they are part of a phrase. Many phrases will also be made up of common words, so simply looking at words will not necessarily raise attention to them. To address this problem, we investigated key word N-Grams – sequences of N words. We looked at several values of N: 2, 3, 4, and 5.

Looking at Bigrams, some pairs of words became apparent that had not been key individually. "Flight times" and "perspective lines" were both overused in FE discussion, and both paint a slightly more vivid picture of what people are talking about. These are examples of discussion relating to proofs of the Earth's shape. An example of a Trigram that lends new insight is "the ice wall". This relates to an idea that the Earth's disc is surrounded by a wall of ice.

4-grams begin to look more like phrases. "Distance to the horizon" and "the Earth is accelerating" are both key in FE discussion. A surprising key 4-gram of the off-topic section was "the state of Israel". Various keywords and phrases relating to Judaism and Israel have appeared throughout analysis. It remains to be seen what context in which it is being discussed, but it would be interesting to see if it links to wider anti-Semitic conspiracy theories [Bilewicz et al., 2013, Kofta et al., 2020]. Other works have found that such conspiracies are prevalent in certain online communities [Allington et al., 2021, Zannettou et al., 2018b].

Table 6.14: Table showing the frequency in each section of the corpus, as well as the Log-Ratio, for the 6 most over-used entity types in both FE and off-topic. Log-ratio suggests how much an entity type was overused in the FE section compared to off-topic.

	QUANTITY	LOC	NORP	MONEY	LANGUAGE	PERSON
Flat Earth	24,498	39,229	5,563	1,061	421	32,798
Off-Topic	1,193	2,506	14,843	1,681	647	44,504
Log-Ratio	3.60	3.21	-2.18	-1.43	-1.38	-1.20

Finally, 5-grams reveal some key phrases and common arguments in Flat Earth discussion. Certain phrases, such as "there is no flat earth" and "if the earth was/were flat", tell us that there are people disagreeing with the theories on the forum. These are the kinds of phrases that could potentially be used to identify groups of Round Earthers. There are other phrases that point towards scientific style discussion, for example "In the flat earth model", and "light travels in straight lines". None of the phrases mentioned appear at all in the off-topic section. This backs up our suggestion from earlier that Flat Earth subject matter is not discussed much in the off-topic boards.

6.6.3 Named Entities

As mentioned in Section 6.3, the tokenisation pipeline we used also performed named entity recognition. This is a process by which all entities in a text are identified. In a similar manner to Section 6.6.1, we looked at entities that were significantly over-mentioned in Flat-Earth posts compared to off-topic.

Table 6.14 shows the frequencies and log-ratio for the 8 most over-used words in either section. As we can see, the FE parts of the forum overused QUANTITY and LOC (location) entities. This backs up the idea that FE debate is very detail oriented compared to regular discussion. Off-topic discussion, meanwhile, contains more NORP (Nationalities or religious/political groups), MONEY, LANGUAGE, and PERSON entities. You may expect these entities to be discussed in off-topic conversation, relating more to topics such as politics.

Various names are highlighted, notably the names of scientists. "Rowbotham" and "Cavendish" are both prominent examples. Samuel Rowbotham was the man who wrote "Zetetic Astronomy: Earth Not a Globe" – the text on which much of Flat Earth theory is built. Cavendish relates to the "Cavendish Experiment" - a famous experiment in which, among other things, the mass of the Earth was calculated.

Key concepts of FET also come up, such as "Universal Acceleration"/"UA" – a popular theory which asserts that the Flat Earth is always accelerating upwards at $9.8m/s^2$, thus explaining the apparent effects of gravity. Useful acronyms are also highlighted such as "ENaG" (Earth Not a Globe). Finding these types of phrases and abbreviations may help in finding terms that distinguish members of the community from newcomers/outsiders.

As with keywords, the off-topic key entities mainly just showed the typical topics of conversation outside of FET. These included the names of games ("Morrowind"), musicians ("(Frank) Zappa"), and politicians ("Biden"). Similarly to the rest of our analysis so far, this suggests that the forum functions as a general social space, where many topics are discussed.

6.6.4 Topic Modelling

To further get an impression of the topics being discussed, we used LDA to identify topics on the forum. The various limitations of LDA have already been discussed in Section 2.1.2, particularly for short texts. We still find it useful here, however, as a way to highlight some possible topics. The topics can be compared to our findings from the rest of this section to build a clearer idea of what type of discussion takes place on FE sections of the forum.

Initially we did this for the entire forum, using 10 as the number of topics[16]. Perhaps predictably, this produced topics that seemed relatively similar to our distinctions of Flat Earth, off-topic, and miscellaneous. Some of the topics related to admin, some to FET, and others more general. While this does confirm that the boards do seem to correspond to the topics discussed on them, it does not give much insight into the key topics of Flat Earth Theory.

To remedy this problem, we used LDA just on the FE boards. This might highlight common areas of discussion or arguments surrounding the Flat Earth. Of the topics produced one contains words relating to "Poles", "Antarctica", and "Australia". All of these words pertain to what is at the edge of the disc. The existence of poles, and ability to travel round the world are often suggested by "Round-Earthers" (RE'ers) as evidence

[16] 10 was chosen arbitrarily, as we did not know how many topics there would be in advance. The number seems small enough that the topics will not be too specific, but large enough that it does not mash everything together.

Table 6.15: Table showing the five nearest neighbours of four example words that demonstrate differences between FE and off-topic sections of the FES forum.

	plot				
Flat-Earth	measure	draw	use	determine	fly
Off-Topic	story	movie	character	album	characters
	wing				
Flat-Earth	air	gas	water	pressure	land
Off-Topic	-	left	media	now	self
	white				
Flat-Earth	bottom	red	black	blue	green
Off-Topic	black	who	kill	jewish	bad
	cancer				
Flat-Earth	degrees	equator	cancer	capricorn	summer
Off-Topic	many	those	jews	these	who

that the Earth is a globe. Another topic appears to relate to experiments and the horizon. The horizon is another topic that is frequently discussed on the forum, as it is often brought up by RE'ers to prove the Earth's curvature. One topic seems to do with the conspiracy side of things. Words such as "NASA", "government", and "conspiracy" appear here. There are also several topics with more varied words, which do not seem as on-brand. It is difficult to understand what these could mean.

Performing LDA on off-topic may help to get an impression of what topics members discuss in the off-topic boards. In line with the keyword analysis, the topics include: health, religion, Donald Trump, video games, and technology. These discussion points are typical of general internet discussion, and further point to the forum acting as a general social space in addition to a platform for debating the Flat Earth.

6.6.5 Word Vectors

As we have already shown in Section 4.4, word embeddings can be a useful tool for looking at language change. In this case, it is not change over time we are interested in, but rather change between FE and off-topic sections of the forum. To look at this, Word2Vec models [Mikolov et al., 2013] were trained on each section of the forum, and the words that changed in meaning the most between these models were looked at, using the method introduced by [Gonen et al., 2020].

Many of the most highly changing words make a lot of sense. For example, the word "casting" was the most different between sections, referring to the casting of light in the

FE section, and entertainment in off-topic.

Table 6.15 shows the neighbouring words for four selected terms that demonstrate differences between the two sections. These words were all selected from the top ten most-changing words, with the condition that they appear at least 100 times in each section. From this, we can see the word "plot" refers to measurements in FE, and fiction in off-topic. The word "wing" seemingly has more aviation-related meaning in FE, versus a more political bent in off-topic. "White" neighbours with colours in FE, while it appears to refer more to race in off-topic. Finally, "cancer" refers to the Tropic of Cancer in the FE section. In off-topic this word confusingly neighbours "Jews". This links to points we have previously made, about Judaism being a subject of a strangely large amount of discussion on the forum.

We also looked at the most closely neighbouring vectors for some of the keywords we identified earlier in the chapter, to shed greater light on the usage of these words, and to find similar ones.. Most of the neighbours we looked at were entirely predictable - for example that "sunrise" neighbours "sunset". Some were slightly more informative. "Rowbotham" had a similar vector to many other famous scientists and philosophers, such as "Einstein" and "Newton", possibly implying that he is discussed on a similar level, and very much considered a scientist.

6.6.6 Parts-of-Speech

The next feature we looked at was Parts of Speech. These are categories of words with similar gramatical properties. Examples of parts of speech are nouns, adjectives, and pronouns. The PoS tags we used were those allocated by Stanza as described in Section 6.3.1. Parts of speech can tell us about the grammar of users in different groups, and the style of the language.

As with words, we began by looking for key parts of speech that are overused in Flat-Earth parts of the forum compared to Off-Topic sections and vice-versa. The results of this showed that Flat-Earth sections of the forum overused numbers, nouns, and symbols. This could be a product of the more technical discussion that seems to take place in these sections based on our previous analysis. Off-topic discussion, meanwhile, overused proper nouns, which may be down to the topics of discussion focusing on politics and entertainment. Pronouns and interjections were also overused,

which possibly points to a more casual writing style in this section of the forum.

We also looked at PoS trigrams, to see if any particular phrase constructions or arguments were overused by any section. These told a similar story, with key FE trigrams containing more numbers, punctuation, and symbols, and key off-topic trigrams containing more interjections, proper nouns, and verbs.

6.6.7 Function Words

Function words are words that contribute to a sentence syntactically, but do not have much of a meaning in themselves. Examples include common words such as “the”, and ‘a’, as well as pronouns, and conjunctions. These words can be useful when looking at author style [Pennebaker, 2013]. This is because authors use these words subconsciously, unlike content words which may be more specifically chosen. In many NLP tasks, these words are removed because they are not useful for many problems, such as topic analysis. We are looking at them here because we are interested in whether or not there are subtle clues as to the style of Flat Earth discussion compared to off-topic.

Many of the key function words for the Flat Earth boards are relating to position or space. Words such as “above”, “around”, “below”, and “beneath” are all key. This could be because of the fact that Flat Earth discussion often involves descriptions of the disc/globe. One thing that is interesting about this is that it suggests that certain linguistic features of deception that have been applied to false information, are simply not applicable to all conspiracy theories. Usually, a lack of spatial vocabulary can be seen as a feature of deception because deceptive texts tend to be less detailed. However, in our case we have found them to be overused in the part of the forum relating to false information. In fact, throughout the study, the Flat Earth parts of the forum appear to be more complex than the rest.

Many of the key function words in the off-topic section are pronouns, suggesting more personal language. It would make sense that this part of the forum was more casual.

6.6.8 Character N-Grams

We have already described word N-Grams, but character N-Grams can also be useful. Character N-Grams capture style because they pick up on sub-word features, e.g.

suffixes. They can also pick up on spelling variations, though this may not be too relevant here.

Looking at character trigrams highlights some shorter sequences of characters that would not be highlighted by word ngram analysis. Acronyms such as "NAG" (Not a Globe) "WGS" (World Geodetic System) are highlighted as overused in flat earth discussion. In off topic conversation, it similarly highlights acronyms, but this time ones not relating to the flat earth, for example "BLM", "RPG", and "GOP". Longer trigrams approach similar results to those for word unigrams. The overused 5-grams, for example, highlight various substrings from within words such as "circumference".

6.6.9 Profanity

Now, we will look at the profanity feature. This was calculated using the *profanity_check*[17] python package, which uses an SVM classifier, trained on 200K human-labelled text samples, to predict profanity. While not perfect, this method will detect many common offensive phrases and terms, but is not limited to those on a restrictive wordlist. For each post, we predict a probability of profanity which indicates how likely a post is to be offensive. We also make a binary prediction as to whether the post is profane.

Examples of profane posts are mostly short, and make use of common swear words such as "shit" or "fuck". Short expressions such as "fuck off" rank particularly highly. Posts with a very low profanity score tend to be long, and fairly dry, posts often involving technical discussion.

Overall, 3.8% of the posts in the forum were judged as profane. Breaking it down by board type shows a large difference, however. 7.9% of posts in the off-topic section are offensive, compared to only 0.9% in the flat earth boards. This is possibly due to off-topic boards containing more casual conversation. Interestingly, it may suggest that the Flat Earth discussion is not full of abuse and arguments. It is worth noting too that the Flat Earth areas of the forum are heavily moderated. This could mean that abusive/profane posts are always removed, which might explain the significant difference. The FE section's level of profanity is lower than `r/science` and `r/conspiracy`, for which the proportion of profane posts are 4% and 10%

[17] `https://pypi.org/project/profanity-check/`

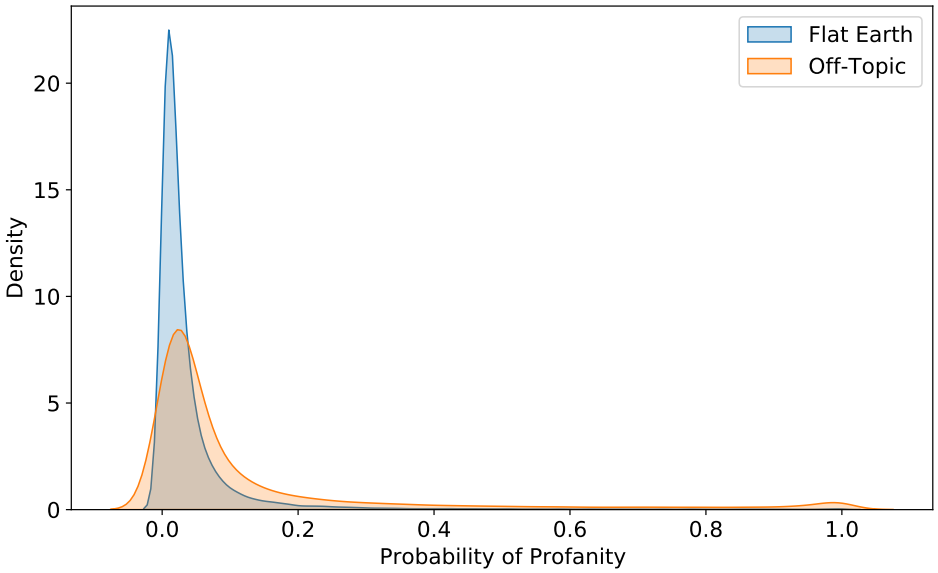


Figure 6.15: Density plot showing distribution of probability of profanity across posts for Flat-Earth boards and off-topic boards.

respectively, suggesting that it is low compared to general internet discussion, while the off-topic section is typically profane.

Figure 6.15 shows the distribution of the probability of profanity in Flat-Earth and off-topic boards. It demonstrates that, while both sections are mostly non-profane, the off-topic section is skewed higher. This difference was statistically significant according to a Mann-Whitney U test ($p < 0.025$).

Figure 6.16 shows the percentage of profane posts in each section (Flat-Earth and Off-Topic) over time. Both seem to go down over time. There is a sustained decrease in the probability of profanity over time in Flat-Earth posts. Could this suggest some kind of strictening of the moderation, or a change in the posts themselves? The Off-Topic section is much more erratic. Though it seems to trend down, there are various peaks throughout. It would be interesting to see whether these corresponded to events or if they were random fluctuations.

Figure 6.17 shows the keywords for profane and non-profane posts on the forum. Unsurprisingly, the profane keywords are mostly common swear words, as one might expect. Interestingly, the words “Jews” and “Israel” both appear as key. This, as with some of the off-topic keywords that came up earlier, suggests a possible layer of antisemitism within the community. Given there are many conspiracy theories

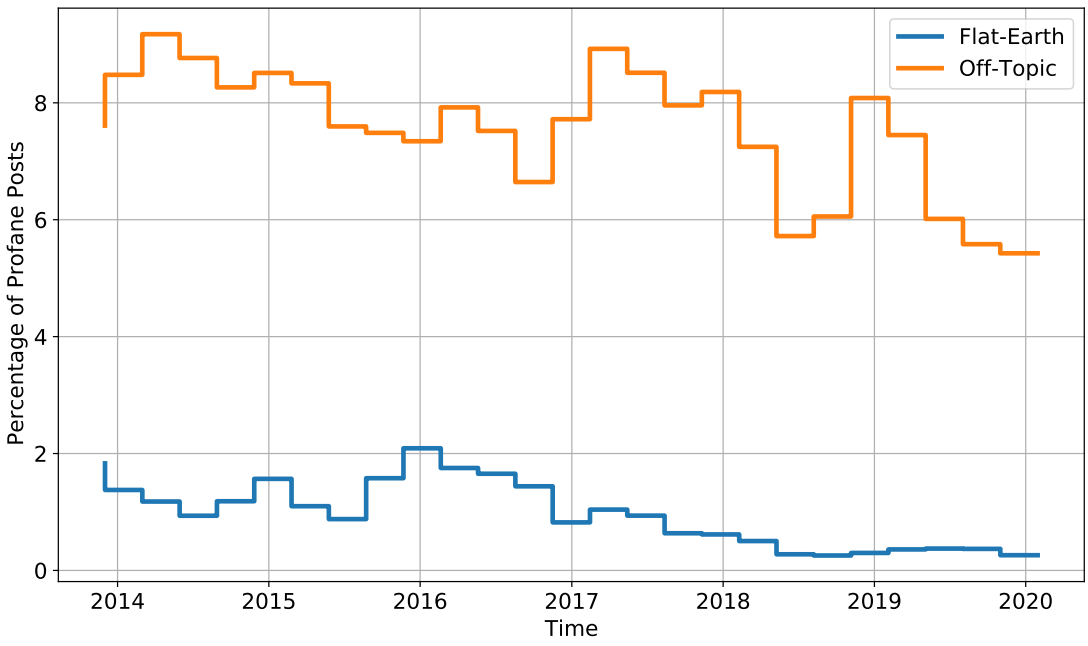


Figure 6.16: Plot of percentage of profane posts over time, using time windows with size 180 days, and step 90 days.

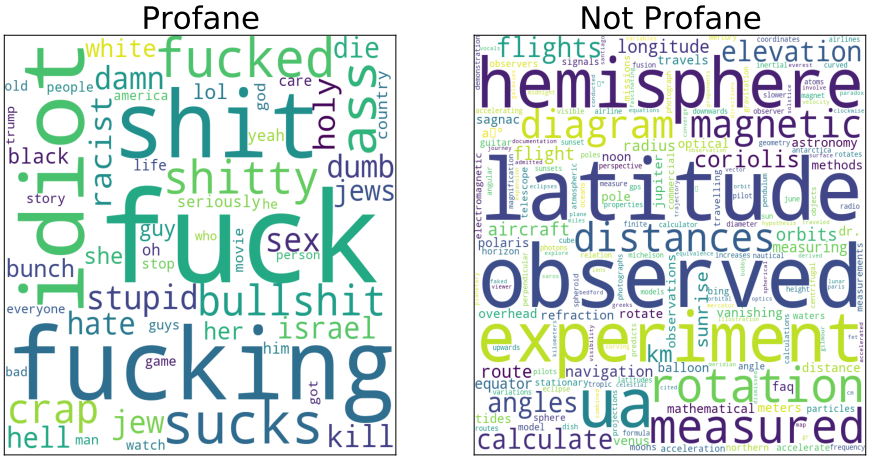


Figure 6.17: Keywords of profane and non-profane posts.

involving Jewish people [Bilewicz et al., 2013, Kofta et al., 2020], which are prevalent online [Allington et al., 2021, Zannettou et al., 2018b], it might be that conspiracy theorists latch onto multiple theories [Goertzel, 1994], particularly when both theories involve secret manipulators, deceiving the world en masse.

The non-profane keywords are largely dry, technical terms. This both suggests that the flat earth related posts are less profane, and also that they are often longer, more technical posts. It is also probably contributed to by the fact that there are far more profane posts in the off-topic section, so to an extent these are keywords of the FE section of the forum.

6.6.10 Summary and Takeaways

In this section, various methods have been used to study the high level language features on a Flat Earth forum. The main outcomes from this have been the identification of words and phrases that appear to define FE discussion over regular conversation. Various technical terms have been collected using data driven techniques. These terms (and phrases) could potentially be used to identify groups of members within the forum. This will allow the comparison of different types of user within these communities.

We also observed, more generally, the notable topics and discussion points of Flat Earth Theory. This has helped to provide a better understanding of what is discussed on the forum. For example, it would seem to suggest that the discussion on the forum is often done in scientific terms, with examples of experiments, etc. The analysis of the off-topic section suggests that the forum is also used as a more general social space by members of the community. Topics discussed by this community also appear to be typical of online fora.

6.7 Discussion

In this chapter we have introduced a new dataset made up of Flat Earth communities, with specific focus on the Flat Earth Society forum. We then performed a meta analysis of these communities, and two comparable non-Flat-Earth communities. Finally, we performed an analysis of the linguistic features of Flat Earth discussion, compared to off-topic discussion in the FES forum. This section will discuss what we have learnt

over the course of our analysis so far, and will help answer RQ3 from Section 1.3.

6.7.1 Meta Features of Flat Earth Communities

In Sections 6.4 and 6.5 we performed a meta-analysis of various Flat Earth communities, as well as two comparable non-Flat-Earth subreddits. Initially, we analysed the posting behaviour in the FES forum.

In this forum we found that there was a small group of prominent users, who made the majority of posts. While most posters only contributed very little (median number of posts was two), these members were very active over the lifetime of the forum, and participated in both Flat Earth related, and more general, discussion. Interesting questions can be asked about the beliefs of these users. For example, are the prominent members all Flat Earth believers, or are some also sceptics?

Another interesting takeaway was that the forum acts as a more general social space, as well as a destination for Flat Earth debate. The second and third most popular boards, after FE debate, were both based around off-topic subjects. However, most visitors to the forum did not post in these areas: ranked by unique visitors, rather than number of posts, these boards were fourth and seventh. This suggests a social dimension to Flat Earth belief, that possibly explains a reason for the community's existence.

The final observation based on the FES forum was how the popularity of the site fluctuated substantially over time. Further analysis should try and understand the external events that may have triggered these shifts in popularity.

Following on from our analysis of the FES forum, we identified several FE subreddits and repeated the meta-analysis on these. The existence of multiple groups suggests that FE discussion is spread across various communities, each possibly having different levels of belief in the Flat Earth, and styles of discussion. Our findings in this meta-analysis were fairly consistent with the forum, with a small group of users dominating.

As with the forum, popularity of the topic seemed to fluctuate. Different communities come and go, with new ones being made continuously. The peaks of popularity were similar, with the subject becoming particularly big in 2018. Interestingly, the subreddits do not become popular until 2016, while the forum had a substantial number of posts as early as 2014. This could point towards the more niche forum being a fermenting ground

for some of the ideas that became more mainstream on Reddit. Zannettou et al. [2017, 2018b] showed how memes and alternative news often originate in small communities, before achieving wider scale influence on mainstream social media. There may be a similar behaviour amongst Flat Earth communities.

We also looked at moderation in the subreddits, by observing the number of removed posts over time. The moderation of different communities seemed to vary from minimal to as high as 60%. Looking at these different FE communities suggests that there are similar posting behaviours across them all, but that they are not completely homogenous.

We also compared these FE communities to two non-FE subreddits: `r/conspiracy` and `r/science`. These groups displayed similar distributions of posts across users, which could suggest that this is more a general feature of online communities than one specific to FE groups. However, we also found that these more mainstream groups had more long-term members. This might suggest that the FE communities have more ephemeral visitors. Finding out why these users come, whether it is for reasons of curiosity or mockery, will be interesting to look into further.

6.7.2 Linguistic Features of Flat Earth Debate

Through our investigations in Section 6.6, we have learned various things about the features of Flat Earth debate. One thing we have observed is the use of heavily technical language throughout the forum. Many of the debates that take place on the forum do attempt to use some form of evidence, often referencing past experiments. Acronyms are common, as are the names of scientists. The community has a range of niche terms, specific to the FE theory which are frequently deployed. If context was removed, one might think one was reading traditional scientific discussion. This is consistent with other conspiracy theories, that justify themselves using pseudo-scientific evidence [van Prooijen and Douglas, 2018] – though Swami et al. [2014] found that, even still, belief in conspiracies was not associated with analytical thinking.

This has interesting comparisons to other false information research which has treated dis/misinformation as deception. Often when people are deceiving they are vague, or use hedging language. In theory, lies are less complex than the truth due to the additional cognitive load of lying [Carlson et al., 2004]. This does not necessarily seem

to be the case here, and it possibly highlights significant flaws in treating disinformation as deception, showing that it should be studied within the context of the specific conspiracy/topic that one is looking at.

A more relevant comparison may be Markowitz and Hancock [2014]'s study of fraudulent scientific literature. They found that fraudulent papers overused scientific terms, possibly in a bid to appear credible. This may be a similar phenomenon to what we have observed in the FE forum.

A deeper dive into this will be necessary to fully understand. Swami et al. [2014] showed that conspiracy believers tended to exhibit lower levels of analytic thinking, while outwardly offering elaborate arguments. To confirm whether or not our findings contradict this, we would need to know the belief of the users exhibiting these features.

In the off-topic sections we identified some keywords involving Judaism. It would be interesting to discover whether this was discussion of religion, or more nefarious references to wider anti-semitic conspiracy theories. Belief in other conspiracy theories has been found to be the number one predictor of belief in another [Goertzel, 1994], so it would not be wholly surprising if it turned out to be the latter.

We also came across an interesting problem with function words, a traditional style feature to look at. Many function words are particularly common in FE debate, such as "round" and "across". This again highlights that many of the features of disinformation within a community may be highly contextual.

Outside of the Flat Earth, many other topics are discussed on the forum. The topics of discussion are exactly what could be expected on any online community: video games, politics, music, etc. This highlights that this community is not simply a place where disinformation foments. It is an active community, in which members engage socially.

6.7.3 Future Work

Building on our analysis, it would be interesting to see if groups could be identified on the fora in an unsupervised way. We know that within these communities there exist individuals with wide ranges of beliefs, and attitudes towards the Flat Earth. By searching for groups, we may be able to learn more about the types of users who reside in these communities.

This leads onto another interesting topic of future work: belief. Understanding the extent to which members of conspiracy communities believe in the theories shared is key to understanding false information on the web. If only a small fraction of members believe, then we are wasting a lot of time and effort on fact checking, as it is somewhat redundant. Belief is almost impossible to determine, however, and studying it would require the labelling of users, which would be time consuming and very difficult without personally interviewing the members in question.

Language change is another area we would like to investigate. For example, it would be interesting to see if the number of trolls or amount of abusive language changed based on FET's popularity in the wider media. It would also be interesting to learn more about how these communities react to events in the outside world, and other, similar communities. We would also be interested to learn what influence, if any, the FES forum exerts on the subreddits. For example, do linguistic innovations appear in the forum before making their way to Reddit?

6.8 Conclusion

In this chapter, we have described the creation of the first dataset of online Flat Earth discussion. This dataset consisted of all posts from one FE forum, as well as eight FE subreddits. A meta-analysis was conducted, to show the posting behaviour of users in these communities. Analysing the forum revealed the use of the site as a general social space, rather than simply a discussion board for FET. This points towards a relatively close-knit community. We also looked at the relationship between the FE forum and external events relating to the Flat Earth. The analysis of the subreddits highlighted the fluctuations of different communities over time, and showed new groups emerging over time.

Following on from this, we performed a linguistic comparison of the FE and off-topic sections of the forum. This revealed features of FE debate. FE discussion contained detailed pseudo-scientific language, while the off-topic discussion was more casual and focused around topics such as politics and entertainment. The analysis also helped us to identify key concepts from Flat Earth theory. Further work will involve more complex linguistic analysis, particularly looking at the change in language over

time and trying to identify sub-groups within the wider community. This is what we turn to in the next chapter.

Chapter 7

Analysing Language Usage in Flat Earth Communities

7.1 Introduction

In Chapter 6, we introduced a new dataset made up of the Flat Earth Society (FES) forum, and Flat Earth (FE) communities from Reddit. We performed a meta analysis of this dataset and looked at the language usage of Flat Earth debate compared to off-topic discussion on the FES forum.

This chapter describes a more detailed analysis of this dataset, focussing on language change and identifying sub-groups of users. The analysis will address the research questions laid out in Section 1.3, primarily seeking to contribute to RQ2 and RQ3. Using a range of analysis techniques, we hope to learn more about the language of Flat Earth communities, and provide insight into online conspiracy communities more generally. We will use the toolbox of methods from Chapter 4 to observe how language usage changes both within and between these communities. Looking at both meta-information and language features, we also wish to identify sub-groups of users within this dataset in an unsupervised fashion. This may give us insight into the types of users that frequent these sites, and the varying levels of belief held by members of the community.

The chapter will be structured as follows. Section 7.2 will look at language change over time on the forum, making use of methods from Chapter 4. These techniques will be used to compare the language of the forum to those of other communities. This will further our understanding of how similar different Flat Earth and related communities

are and how this changes over time, and will help us to answer RQ3. It will also provide an opportunity to test methods from Chapter 4, which will help answer RQ2. This will be followed up in Section 7.3, where we look for logical groupings based on this meta-analysis. Finally, in Section 7.4, we will find clusters of users using some of these linguistic features. Identifying sub-groups within the community will increase our understanding of the types of users, and styles of discussion, on the forum, and will help answer RQ3.

7.2 How does the language of the community change over time?

When the subreddit r/flatearthsociety closed its doors in 2017, "round-earther trolls" were pointed to by the administrator as the primary reason for its closure. In a closing post, the administrator of the group lamented that the community had become a Q&A forum for angry Round Earthers who had recently read about Flat Earth theory online. It would be interesting to try and assess how true this assertion was. Had the forum really been taken over by trolls and, if so, how does this change manifest in the language of the forum? In this section, we will investigate the way language changes over time within the flat earth community. This will increase our understanding of how various Flat Earth communities relate to each other over time, and how language use changes within communities.

We look at three aspects of language change in this section:

1. Looking at changing word usage.

2. Splitting the forum into stages.

3. Comparing Flat Earth communities over time.

Studying these aspects will involve applying various methods from Chapter 4. This analysis will both contribute to answering RQ3 from Section 1.3, and also help us to ascertain the usefulness of the methods on a new dataset, helping to answer RQ2.

7.2.1 Changing Word Usage

In this section, we wish to look for changes in language usage. This means looking at the way that words change over time, in terms of usage. A couple of methods from Chapter 4 will come in useful for this task, namely diachronic word embeddings (Section 4.4), and UFA (Section 4.6).

To begin with, we will look for the word vectors that changed the most over the life of the FES forum. We did this using the method described by Gonen et al. [2020], and demonstrated on Hansard in Section 4.4. This technique involves training word embedding models on each corpus, and comparing the overlap in the neighbours of each word in both corpora.

The general process we used was more or less the same as that used in Section 4.4, although we used windows based on posts instead of time:

1. Split the corpus into windows of 10,000 posts, producing six windows.
2. Train a Word2Vec model [Mikolov et al., 2013] on each window.
3. For each word:
    (a) Calculate the 1,000 nearest neighbours for each time window. Neighbours were found using cosine similarity.
    (b) Get the intersection of the list of nearest neighbours, for each subsequent window.

This will allow us to see which words changed in usage the most at each time window, compared to the last. The words with the greatest change may give us some impression of language change on the forum; for example, emerging topics.

Nearest Neighbours Over Time

The first step of this method was to find the nearest neighbours of a given word at multiple time points. This can be calculated using cosine similarity, because vectors that are spatially near to a word's vector can be considered semantically similar. As an initial exploration, we looked at the nearest neighbours for some strongly FE-related terms.

Table 7.1: Table showing the nearest neighbours for the word "flat" over time.

Window	Neighbours					
01/12/2013	round	globe	shape	evidence	this	theory
	i	fe	believe	model	map	what
30/12/2015	round	globe	shape	fe	evidence	sphere
	spherical	map	proof	believe	wrong	me
16/01/2017	round	globe	fe	shape	map	model
	evidence	believe	real	there	proof	true
09/11/2017	round	globe	fe	shape	believe	map
	spherical	model	evidence	conspiracy	real	theory
20/04/2018	round	globe	fe	model	map	spherical
	shape	sphere	moon	theory	believe	evidence
07/10/2018	round	globe	fe	spherical	map	model
	shape	sphere	believe	i	moon	wrong

For example, Table 7.1 shows the nearest neighbours of the word "flat" over time. As one can see, this word appears to be fairly stable in its usage over time, at least based on the top 12 neighbours. This is probably to be expected in a Flat Earth community. Words such as "round" and "globe" are unsurprisingly similar in usage. Other terms such as "evidence", "believe", and "conspiracy" suggest more about the more conspiratorial aspects of the community.

We found similar results for other Flat Earth related words, such as "earth", "globe", "disc", and "UA" (Universal Acceleration). One interesting observation for the word "disc", was the introduction of "ice" as a neighbour in the third window. This word sticks out as it does not intuitively relate to the word disc. Often in the Flat Earth community, ice is mentioned referring to the "Ice Wall", which some suggest surrounds the disc. Its introduction as a near neighbour may indicate this concept growing more prevalent.

Looking at the neighbours of the word "ice" itself, we observe similar behaviour. From the second window onwards, "wall" becomes a very close neighbour. This could suggest that at some point during the second window, this concept of an ice wall is introduced into mainstream discussion.

We looked at the neighbours over time of the keywords of the Flat Earth boards (compared to the off-topic boards, as decribed in Section 6.6). To a manual inspection, these words also appear to be largely stable. This is not surprising, as one would expect most core concepts of Flat Earth Theory to be relatively consistent over time.

Table 7.2: Table showing the words with the most change on the Flat Earth boards of the FES forum for each pair of consecutive windows.

Time Windows	Ten Most Changing Words				
2013/12/01 to 2015/12/30	corrected slightest	cancer google	respect closed	f expected	alt experts
2015/12/30 to 2017/01/16	particularly became	3d giving	terrible fully	super display	explaining pilots
2017/01/16 to 2017/11/09	parallax powers	compute necessarily	super click	3d particularly	escape rockets
2017/11/09 to 2018/04/20	stream fully	inclined derive	immediately define	club constantly	fits adding
2018/04/20 to 2018/10/07	beat constantly	fits finally	assumed furthermore	closely stops	seismic wires

Finding the Words With the Most Change

Once we had calculated the nearest neighbours, we could look at how much they changed between each window. To do this we use the measure proposed by Gonen et al. [2020], using the intersections of the word's top 1000 nearest neighbours between each consecutive pair of windows. The lower this intersection is, the more a word has changed between windows. By doing this, we can identify the n most changing words.

One issue with this approach is that it will naturally pick out words that appear in a given window for the first time. This is potentially useful for observing new terms, but not at all for looking at the change in usage of existing ones. For this reason, we restricted the words that could be highlighted to ones that appear in all windows. We also filtered out punctuation because it does not tell us anything useful.

Table 7.2 shows the ten most changing words for each pair of time windows. It is hard to interpret what these findings mean. The words by themselves mean nothing intuitively, even when looking at the nearest neighbours at each window.

To limit the terms to specific ones of interest, we filtered the output down to only words which are key in Flat Earth boards. Even after doing this, the results remained difficult to understand. Looking manually at the neighbours of selected words over time did not make them any clearer.

For the most part, the words highlighted by this technique are unintuitive. One potential reason for this is that the corpus is not large enough to train good word embeddings. This is especially likely once it has been split into smaller windows. It is, however, also possible that there simply is not much change over a such a relatively

short span of time.

The solution to the problem of corpus size may be to use a pretrained Word2Vec model, and continue training it on the Flat Earth corpus. This would provide better quality embeddings, and would mean that rarer words could still have had enough examples to train embeddings. We tried this using a Word2Vec model trained on the `r/science` subreddit, but we found that it simply meant that the same FE specific words came up as highly changing. For example, the word "flat" unsurprisingly takes on a new meaning on the FES forum, compared to `r/science`.

7.2.2 Collocates Over Time

Following on from the word embedding analysis, we will now look at collocates of some notable words, and see if this suggests any shift in meaning. We will use UFA to plot the usage of terms over time, as we did in Section 4.6. Collocation is frequently analysed in corpus linguistics, and involves looking for words that co-occur together more commonly than one might expect by chance. This method has a similar motivation to looking at neighbouring word vectors, but uses a much more simple process.

To begin with, we looked at the collocates of some common FE words. As with the diachronic embeddings, we see that most of these words are fairly stable, both based on their UFA plots and manually looking at their collocates over time. There are, however, a couple of words that did appear to undergo some change.

The first of these we will discuss is the word "ice". We have already mentioned how this may have changed in the previous section. Plotting its collocates appears to demonstrate some interesting changes. Figure 7.1 shows two collocates of the word "ice": "wall" and "ring". The vast majority of "wall" mentions co-occur with "ice" throughout the time period, and the frequency of the word increases up to a peak in 2016. This suggests that the topic "ice wall" is increasingly popular, but is stable in its meaning over time. The proportion of occurrences of the word "ring", on the other hand, decrease over time. This could demonstrate "ice wall" becoming the dominant way of describing the wall/ring of ice that surrounds the Earth's plane in many FE models.

Another FE term we found changed in an interesting way was the word "round". Figure 7.2 shows the proportion of occurrences that were co-occurances for four collocates of "round". The plot gives some idea of how the usage of the word "round"

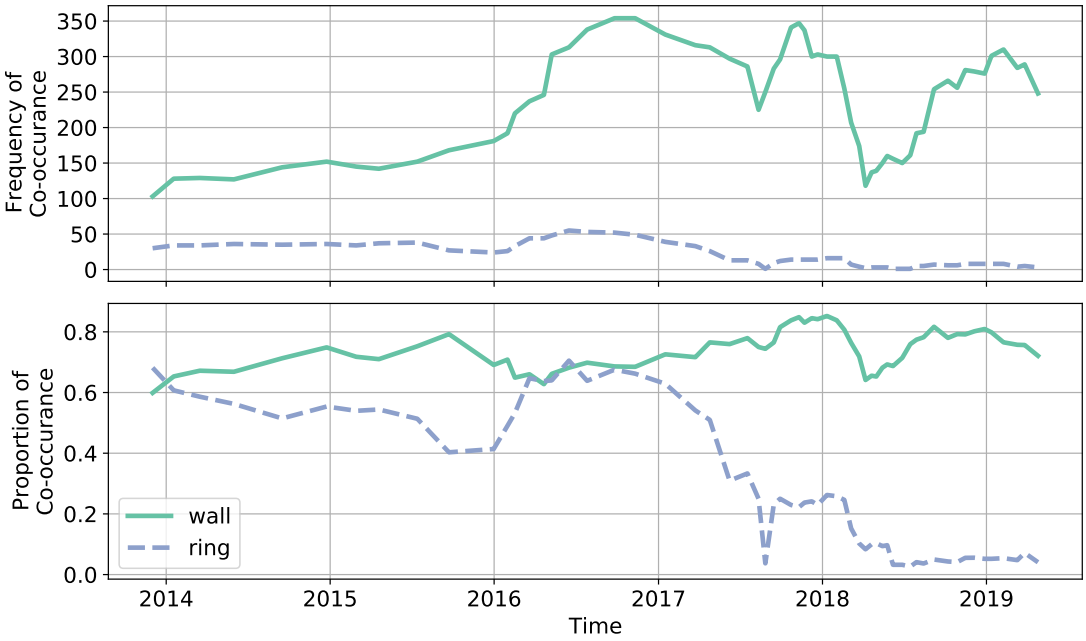


Figure 7.1: Plot showing two collocates of the word "ice" on the FE boards of the FES forum. Top plot shows the number of co-occurrences between each word and "ice", and the bottom shows the proportion of each word's occurrences that were co-occurrences.

fluctuates on the forum. Its peak in late 2017 suggests that more people were using the word round to refer to the round earth, and most dramatically "round Earthers" ("RE'ers"). This could suggest more RE'ers on the forum, or at the very least more mention of them. This could line up with the idea, mentioned at the beginning of the chapter, that RE'ers took over FE communities.

Figure 7.3 shows the UFA plot for the word "round". There is a trough in mid-2017 and peak in 2018 that suggest the collocates become more consistent after 2017. This period of change corresponds to the fall in co-occurences between "round" and the terms shown in Figure 7.2. It is possible that the usage of "round" stabilises, but not around the terms shown in the figure. This could possibly indicate that the flood of round Earthers has subsided during this period.

The analysis in this section has not found a huge amount of usage change on the forum. We have, however, highlighted a couple of examples. Despite these examples, it seems that word usage on the forum is fairly stable over time, though it is possible we simply do not have enough data to detect meaningful change.

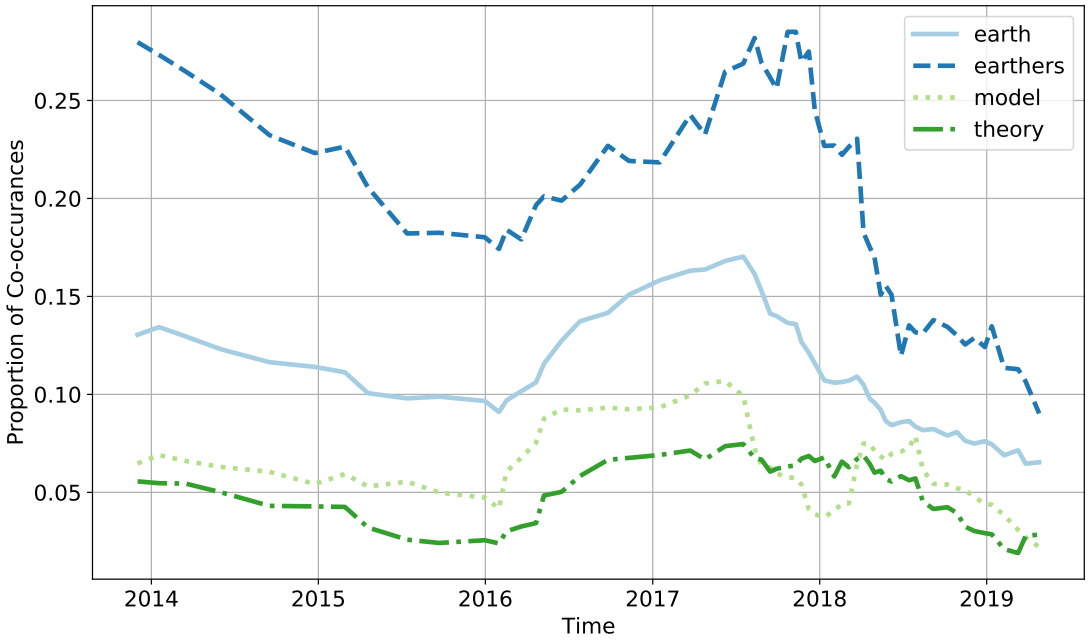


Figure 7.2: Plot showing the four collocates of the word "round" on the FE boards of the FES forum. Shows the proportion of each words occurrences that were co-occurrences.

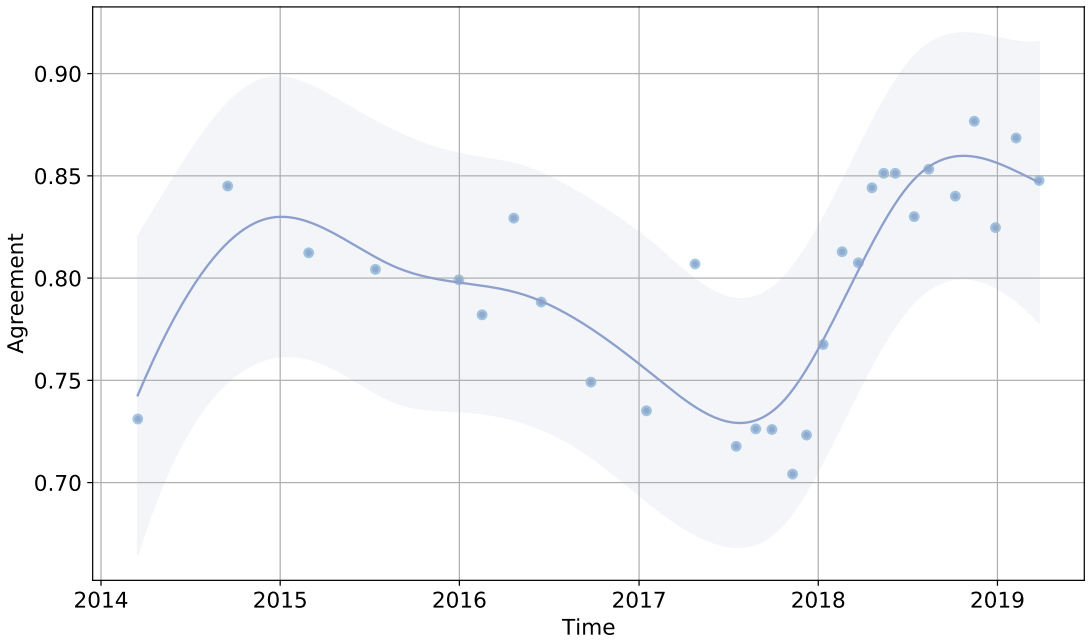


Figure 7.3: UFA plot showing the fluctuation of collocates over time for the word "round".

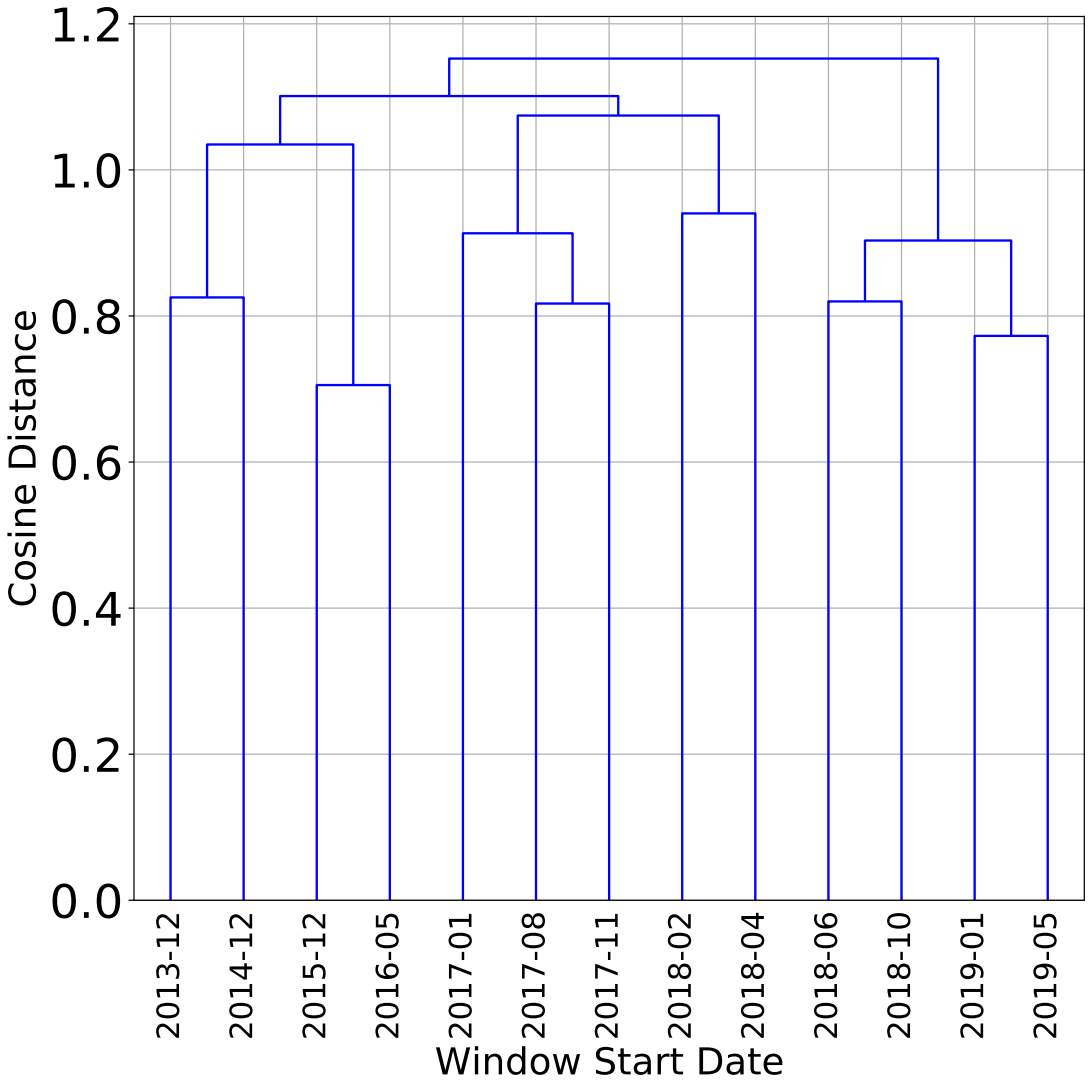


Figure 7.4: VNC plot showing stages of the Flat-Earth-related boards on the FES forum. The relative frequencies of the top 1000 words were used as features. Frequencies were standardised by removing the mean and scaling to unit variance. Windows consisted of 5000 posts. Window start dates were rounded down to month for ease of reading.

7.2.3 Identifying Stages

So far we have examined usage change in the FE community by looking for words that change in their meaning over time. The next thing we are interested in doing is identifying linguistic stages within the forum. To do this, we will use Variability-based Neighbour Clustering (VNC), which has already been described in Section 4.5. By analysing these stages and comparing them to other plots, we can learn more about the changing characteristics of the forum over time. For example, we may observe periods where different topics are popular, or where the language of the forum changes.

Figure 7.4 shows the VNC plot for time windows in the FES forum. If we choose a

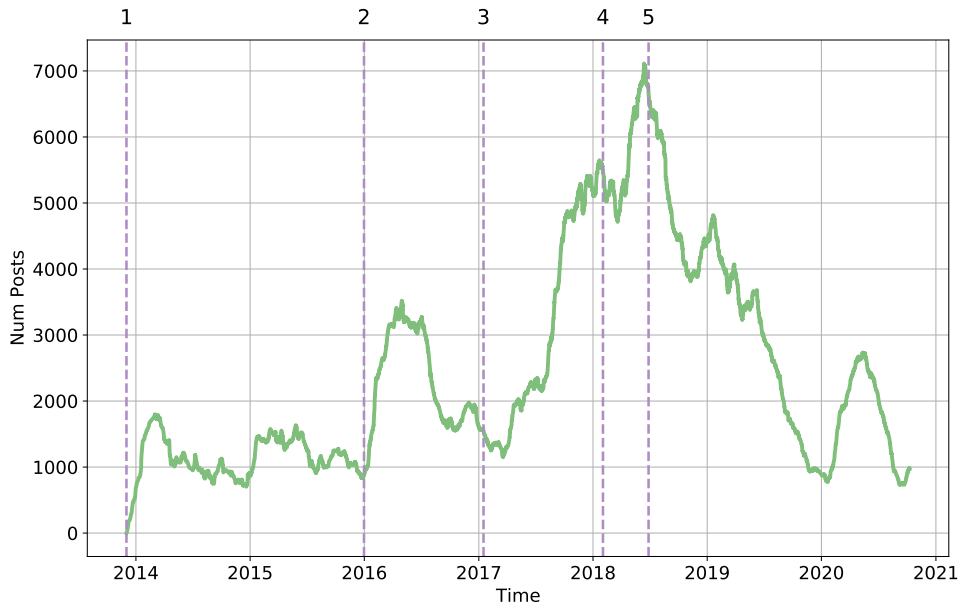


Figure 7.5: Plot showing the beginning of each stage according to the VNC showed in Figure 7.4 with a cut-off of 1. This is shown against the number of FE posts over time, using a rolling window of 90 days.

cut-off of 1, this gives us five stages. For easier comparison, we plot the beginnings of each of these windows against the number of posts over time, in Figure 7.5. At a glance, the stages appear to correspond relatively logically to early periods in the meta-graph.

The first of these stages seems to correspond to the early period of the forum, leading up to its initial rise in posting. The second follows on from this during a period which shows a decline in the number of posts. The third stage takes place during a large increase in posting on the forum, nearly up to its global peak. This seems to be the height of popularity for the forum. The fourth window contains the tip of this peak, and the beginning of the decline, while the fifth follows this decline in popularity all the way down to pre-peak levels.

We can observe interesting things about these stages by looking at the keywords for each[1]. For example, in the third stage (during the large increase in posts), "Shaq" is a keyword. This refers to American former professional basketball player, Shaquille O'Neal, "Shaq" for short. O'Neal said in an interview in 2017 that he believed the

[1]Words overused by posts in a given stage, compared to all others. Keywords were found using Log-Ratio as in previous sections.

Earth was flat, becoming one of the most high-profile Flat-Earthers[2]. This news story may have raised the profile of the Flat Earth conspiracy, and possibly drove people to the forum, although it is also possible that Shaq himself brought up the topic because it had already begun to become popular.

The third stage also contains the keyword "FE-ers". This word enjoys a substantial increase in frequency during this period. This term could imply that the author is not a Flat-Earther, and that they either oppose the belief, or are an outsider asking questions about it. This would go some way to suggest that the language of the forum during this stage becomes more influenced by users who do not identify as Flat-Earthers. This would make sense in the context of the large increase in posts during this period.

There are some interesting early keywords too. Stage 1 contains the keywords "council" and "vote". These refer to the "Zetetic Council" which occupies a high status on the forum. The number of mentions of "council" drops significantly after this first window. This could suggest that the council was only notable early in the forum's life. It possibly points to membership widening from an initial small community to begin with.

Variants of the word "aether" ("ether", "aether", "aetheric", etc) are also key in the first stage. This word declines in frequency substantially after the early stages of the forum, though it peaks again towards the end. This could be an example of a concept that comes in and out of fashion. Or it could be an example of a term that new users are less familiar with, so it declines in usage during the phase that the forum is most popular.

As well as looking at VNC for bag-of-words, we also looked for stages using part-of-speech trigrams and character trigrams. Both of these feature sets produced almost identical stages to words. The only difference was, for both, that the fourth stage was split between the third and fifth, meaning that there was one stage encompassing the rise in posting, and another containing the fall. Both these other feature sets produced better defined clusters than Bag-of-Words, with PoS Trigrams doing best. This is probably due to these features being less sparse than simple BoW. However, despite this, the similarity of the stages produced by each feature set reassures us that they are somewhat meaningful.

[2]Though he later insisted it was a joke. (`https://tinyurl.com/shaq-clarification`)

Limitations of VNC

There are several key limitations of VNC used in this context, and while we do not believe they make the method useless for our purposes, they are worth bearing in mind. Firstly, the quality of the clusters seems low. The Cophenetic Correlation Coefficient [Sokal and Rohlf, 1962] sits at around 0.78 for BoW, and only increases to 0.85 for PoS Trigrams. The clusters are also formed at rather high difference levels. This again suggests that for windows to be clustered together, they do not need to be hugely similar. However, it is still the case that any windows clustered together are the most similar of any pair of neighbouring clusters, so we believe it is still useful.

Flat Earth Subreddits

After analysing the VNC results of the Flat Earth Society forum, we next looked at the Flat-Earth subreddits introduced in Section 6.5. Figure 7.6 shows the stages produced using VNC for each subreddit, using a cut-off based on each group's dendrogram. Many of the subreddits do not cluster especially well, but by looking at the dendrograms and the keywords of each cluster, we can better understand them.

Some of the subreddits did form possibly meaningful stages. The subreddit `r/flatearth` split into three stages, one taking it up to its peak in 2018, another until early 2019, and then a final stage filling the remaining time. The main peak in posting lines up with the peak in the forum. The first stage contains the period building up to this peak. As with the forum, there was a second smaller peak in early 2019, and another in early 2020. These are not as clearly associated with changes in stage, however.

`r/flatearthsociety` splits into three stages: one leading up to the first peak, a second from this peak to the beginning of the later increase, and a third running until the subreddit's closure. The keywords of these stages do not shed a huge amount of light, but they do provide some hints. "Deleted" is a keyword in the central window, when posting hits its peak. This suggests a correlation between heavier posting and stricter moderation. The keywords in all windows make sense, and seem to broadly stick to FE topics. "Antarctica" and "pole" are both key in the final window. In the forum, we speculated that mentions of this topic may come from new users who have read about that aspect of FET online and come to ask questions. This would fit with the narrative proposed by the subreddit for its closure. Even so, despite "shit" and "stupid" both

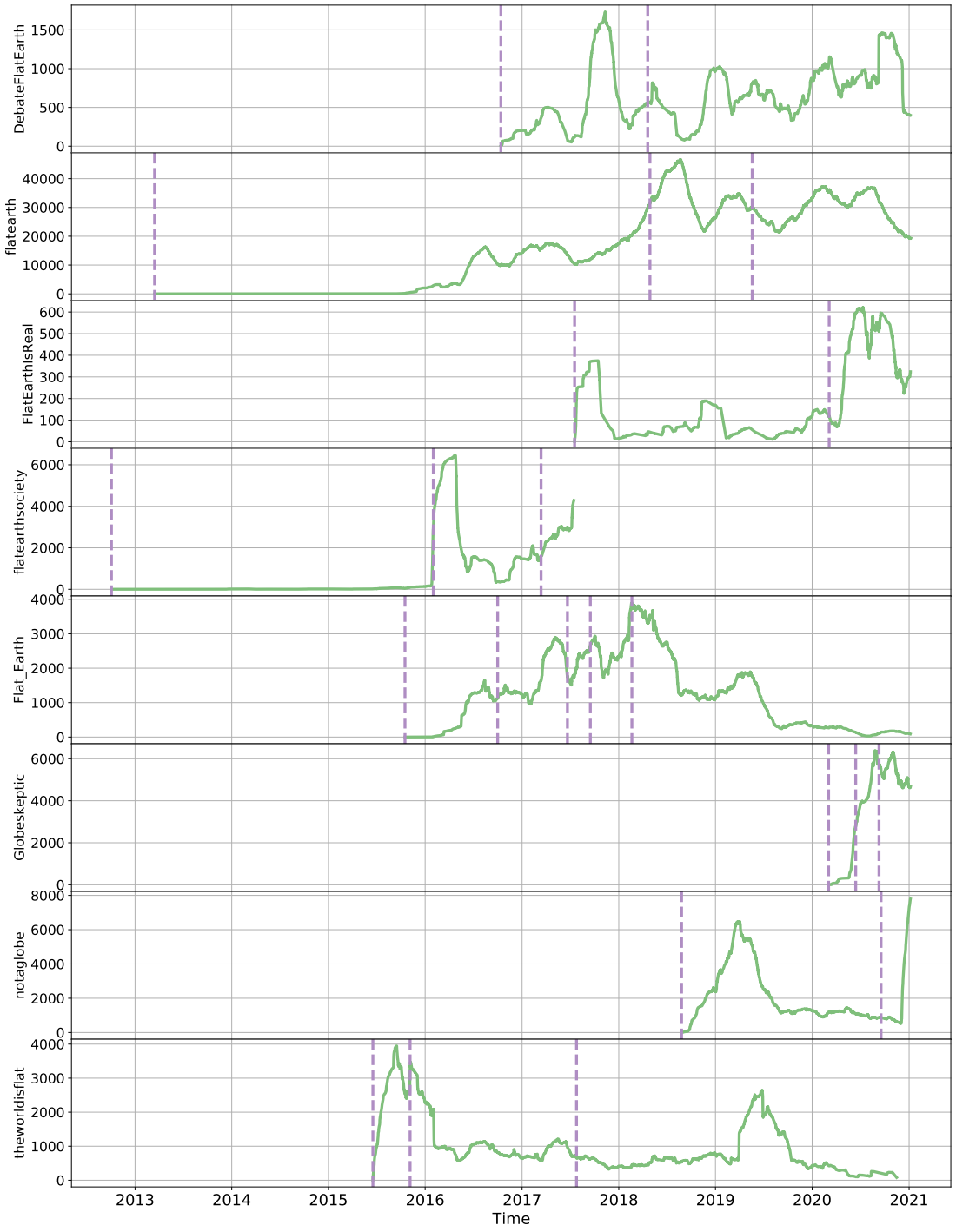


Figure 7.6: Plot showing the beginning of each stage according to the VNC of each FE subreddit with a cut-off chosen for each one based on the dendrogram. This is shown against the number of posts over time, using a rolling window of a set number of posts, based on the number of posts in each subreddit.

appearing as key in the final window, the keywords do not generally seem more abusive, although it is possible offensive comments would have been removed by moderators.

The subreddit r/notaglobe undergoes an interesting change. It was split into two stages: one featuring a peak of posting activity in 2019, and another containing a more recent upswell in 2020. Interestingly, the first stage is defined by many offensive keywords (e.g. "retarded", "troll", "dumb"), while the second contains more technical discussion. This almost suggests the opposite effect than the one suggested as the reason for r/flatearthsociety's closure, and shows that not all of these communities behave the same way. Looking at the migration between these communities would be an interesting subject of future study.

7.2.4 Comparing Language Between Communities

To conclude this section on language change, we will compare the language of various FE communities over time. This analysis will allow us to better understand how similar Flat Earth communities are in their language, as well as seeing how communities diverge and converge over time. To do this, we will calculate the cross-entropy of text from our FE subreddits, compared to a language model trained on the FE boards of the FES forum. We will also compare the language of our two non-FE subreddits, as a point of comparison.

We discussed in Chapter 4 that Cross-Entropy has problems in settings where a small pool of users dominate the discussion. This was a big problem when comparing groups within a community, but not as much of one when we are comparing entire communities to one another. If individuals are very prevalent in a given community, then it does not matter so much that they disproportionately define the language. Therefore, we have chosen to use Cross-Entropy for this task, without the user-based sampling of ACE.

For each text, the cross-entropy is calculated with respect to a snapshot bigram model, trained on text from the current window of the forum. We used bigram models because they are simple and fast to train. By plotting the average cross-entropy for each snapshot, we can get an idea of how the posts diverge or converge with the FES forum over time.

Instead of using full posts as texts, we used regularly sized chunks of text. This is because longer texts will tend towards having a larger cross-entropy, irrespective of the

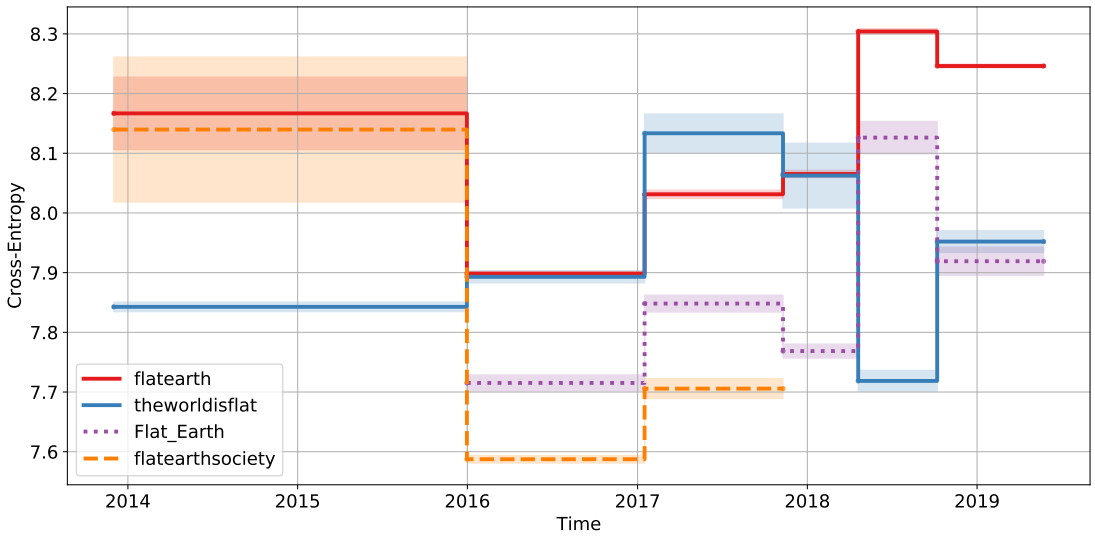


Figure 7.7: Plot showing the cross-entropy per FES forum snapshot model for post-chunks from select Flat-Earth subreddits. Snapshots were trained every 10,000 posts, with a step of 10,000 posts.

similarity of the text to the model. To account for this, each post was split into 30 word chunks, and chunks of less than 30 words were discarded. This meant throwing away some text, but it was preferable to using only the first 30 words of each post, which would have excluded an enormous amount of text.

Figure 7.7 shows the average cross-entropies of the text chunks of various subreddits, for each snapshot model. To assess the stability of the cross-entropies over time, we repeatedly sampled half of the cross-entropies and then plotted the standard deviation of them for each window.

We can observe some interesting behaviours. The first thing worth addressing is the first window, for which `r/flatearth` and `r/flatearthsociety` have strangely high values of cross-entropy. This may be because, at this point in time, there are very few posts on either subreddit. Neither has its increase in posting until 2016, when both appear to stabilise. `r/theworldisflat`, on the other hand, has more posts at this early stage in time, and begins with a CE more in line with future windows.

Both `r/flatearth` and `r/Flat_Earth` have a spike in cross-entropy in mid 2018. Both of these subreddits also experience a drop in number of posts at this point. It is unclear why a reduction in the number of posts would correlate to an increase in cross-entropy, but as it is an average, there is no particular reason why CE would increase as posts decrease, especially without the standard deviation becoming much

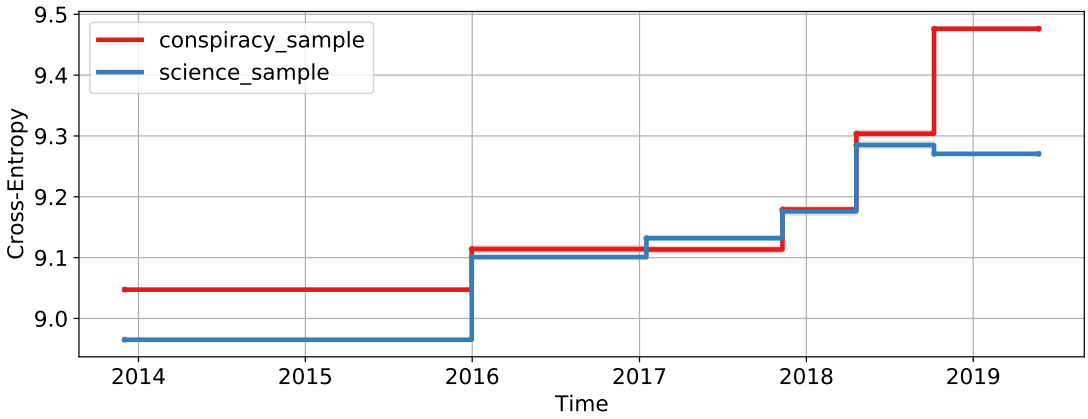


Figure 7.8: Plot showing the average cross-entropy per FES forum snapshot model for post-chunks from `r/conspiracy` and `r/science`. Snapshots were trained every 10,000 posts, with a step of 10,000 posts.

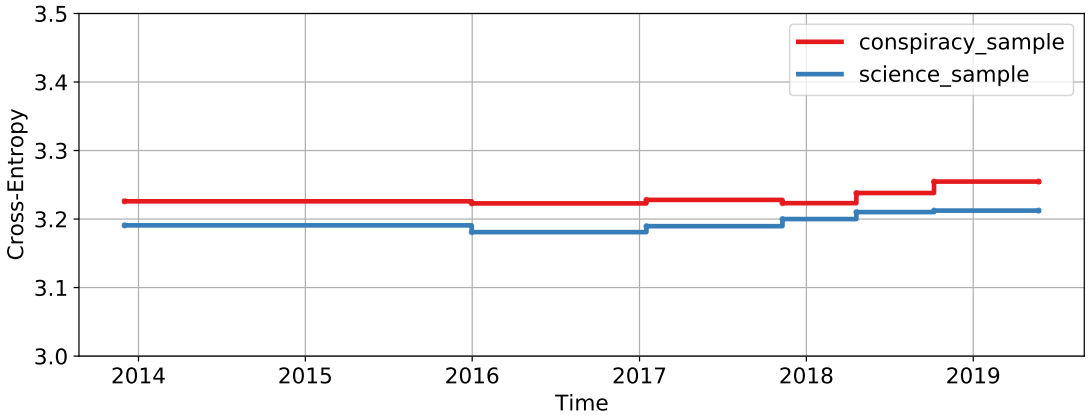


Figure 7.9: Plot showing the average cross-entropy per FES forum snapshot model for post-chunks from `r/conspiracy` and `r/science`, using parts-of-speech rather than words. Snapshots were trained every 10,000 posts, with a step of 10,000 posts.

larger.

`r/flatearthsociety` has the lowest CE relative to the forum over time. This is what we would hope to see, as it is affiliated with the forum, and possibly shares users and style. Its CE does increase in the final stage of its life, possibly giving credence to the notion that it had an influx of new members. However, this increase is not as substantial as for some of the other communities.

`r/theworldisflat` and `r/flatearth` follow a similar trajectory of becoming more divergent over time. That is, until 2018, when the CE of `r/theworldisflat` drops substantially, while `r/flatearth` increases substantially. No notable change in `r/theworldisflat`'s posting behaviour happens during this period.

Figure 7.8 shows the average CE of post chunks from the r/conspiracy and r/science subreddits. The CE values for these two subreddits are still much higher than any of the FE subreddits, which is what one would expect to see. Both of these subreddits grow more divergent from the FES forum over time. They seem to do this in a similar manner, although r/conspiracy begins with a higher entropy, and also grows suddenly much more divergent in the final window. It is worth bearing in mind, though, that this is using words as features, and because of this will be more influenced by topic than style. It may just be that the usage of scientific vocabulary on the FES forum makes it that little bit less different to r/science. When looking at parts-of-speech instead of words, as shown in Figure 7.9, we found that r/conspiracy was slightly more divergent from the FES forum than r/science, suggesting that the language style of tfes.org was more similar to r/science than r/conspiracy.

It is interesting that the FES forum is less divergent from the science than conspiracy subreddit. This could suggest that the scientific nature of the discussion is more important to the community's language than the fact they are discussing conspiracy theories. It may also indicate that the language of different conspiracies is not homogenous, and that false information research must think carefully about the specific domains being studied.

7.2.5 Language Change Summary

Over the course of this section, we have performed various analyses of diachronic language change on the forum. Our findings have not found any dramatic shifts, particularly in word usage which seems relatively stable. Some results have pointed towards the alleged increase in "Round-Earthers" in these communities over time, but further analysis will be needed to better understand changes in levels of belief within the FE community.

7.3 Finding Meta-Groups

In Section 6.4, we demonstrated that the vast majority of users on flat earth fora post very little. Many users (almost half) only posted once, and similarly most users posted only on one board and were only active for a single day. The majority of posts are

made by a relatively small subset of highly active users. We would like to investigate whether discussion on the forum is led by the ephemeral visitors, who make up the vast majority of members, or by the core-community of the forum, who comprise the majority of posts. More generally, we aim to see if we can identify groups of users within the community, based on their posting behaviour. This will help to answer RQ3 from Section 1.3, by increasing our understanding of FE communities.

7.3.1 Identifying Groups of Interest

In Section 6.4, a general meta-analysis of the Flat Earth Society forum was performed. Here, analysis will use some of the observations from that section to split the membership of the forum into meta-groups.

Building on the meta-analysis, we split the forum into groups of users using k-means clustering. Each user was represented by three statistics:

1. The number of posts they contributed.
2. The number of boards they posted in.
3. The number of days they were active on the forum.

These features were all logged, so they are more evenly distributed, and to reduce the significance of larger differences in count. Next the logged values were scaled so that the maximum value is one, and the minimum is zero.

K-means clustering was chosen due to its simplicity and ubiquity. We also experimented with other clustering methods, such as mean-shift, and the results were not substantially different. Three was chosen as the number of clusters. This value was selected using the elbow method [Yuan and Yang, 2019].

The k-means clustering produced three groups, which we have named **Ephemeral Visitors**, **The Middle**, and **Core Community**. Figure 7.10 shows these clusters. There

Table 7.3: Table showing the median value of each meta-feature for each of the K-means clustered groups.

	Num Posts	**Num Boards**	**Num Days**
Ephemeral Visitors	1	1	1
The Middle	6	1	11
Core Community	52	3	246

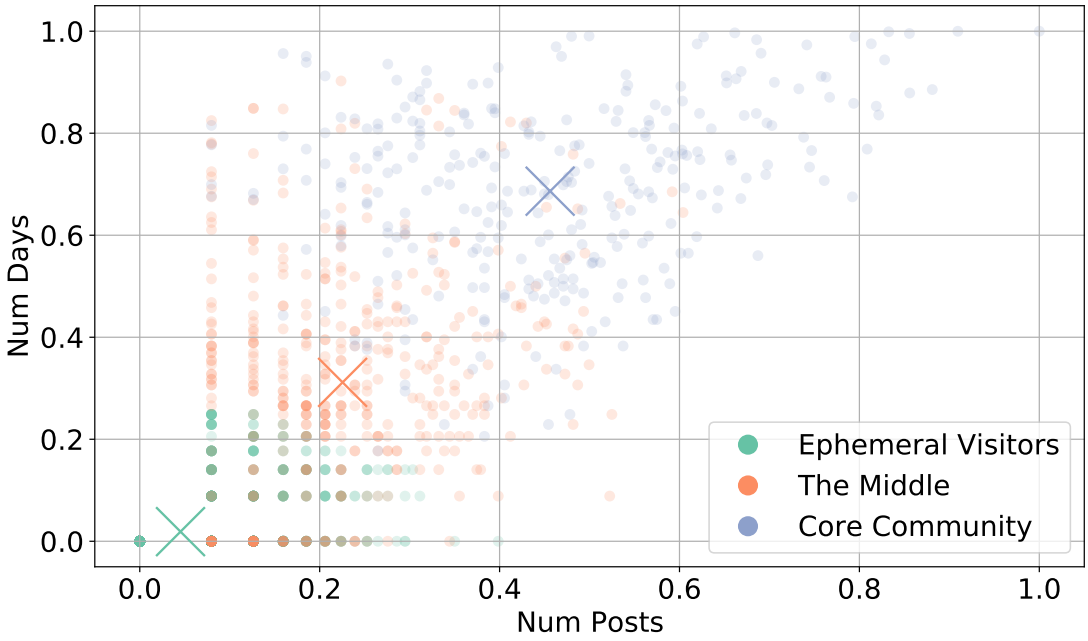


Figure 7.10: Scatter plot showing the meta-groups in the Flat Earth Society forum, based on k-means clustering. Features are scaled between 0 and 1. Opacity represents number of users at a given point.

is no completely clean separation between these three groups, and there is some overlap. The medians of each cluster (Table 7.3), however, suggest each represents users with different characteristics in a way which is useful for analysis. Figure 7.11 shows the distribution of these features for each of the three clusters. Finally, Figure 7.12 shows the number of posts (logged) over time for each of the three groups.

Ephemeral Visitors each contributed very few posts, with most contributing one, and all only posting on a single board. They also, for the most part, were only active for a single day (0 on the scaled plot). This suggests that these users mostly visited the forum to make a single post, and possibly responded to a couple of replies.

Users in **The Middle** also do not contribute many posts, with a median of six, but as a group they have more users that post across different boards, though the median is still a single board. There is also greater spread in the number of posts, with the most prolific member of this group making 191 posts. These users also have a large spread of lifetimes, with some users being active for much of the forum's existence. This suggests that the Middle contains a mix of less active members of the community, and more active visitors. As evident in Figure 7.11, this group has a degree of overlap with the two other groups.

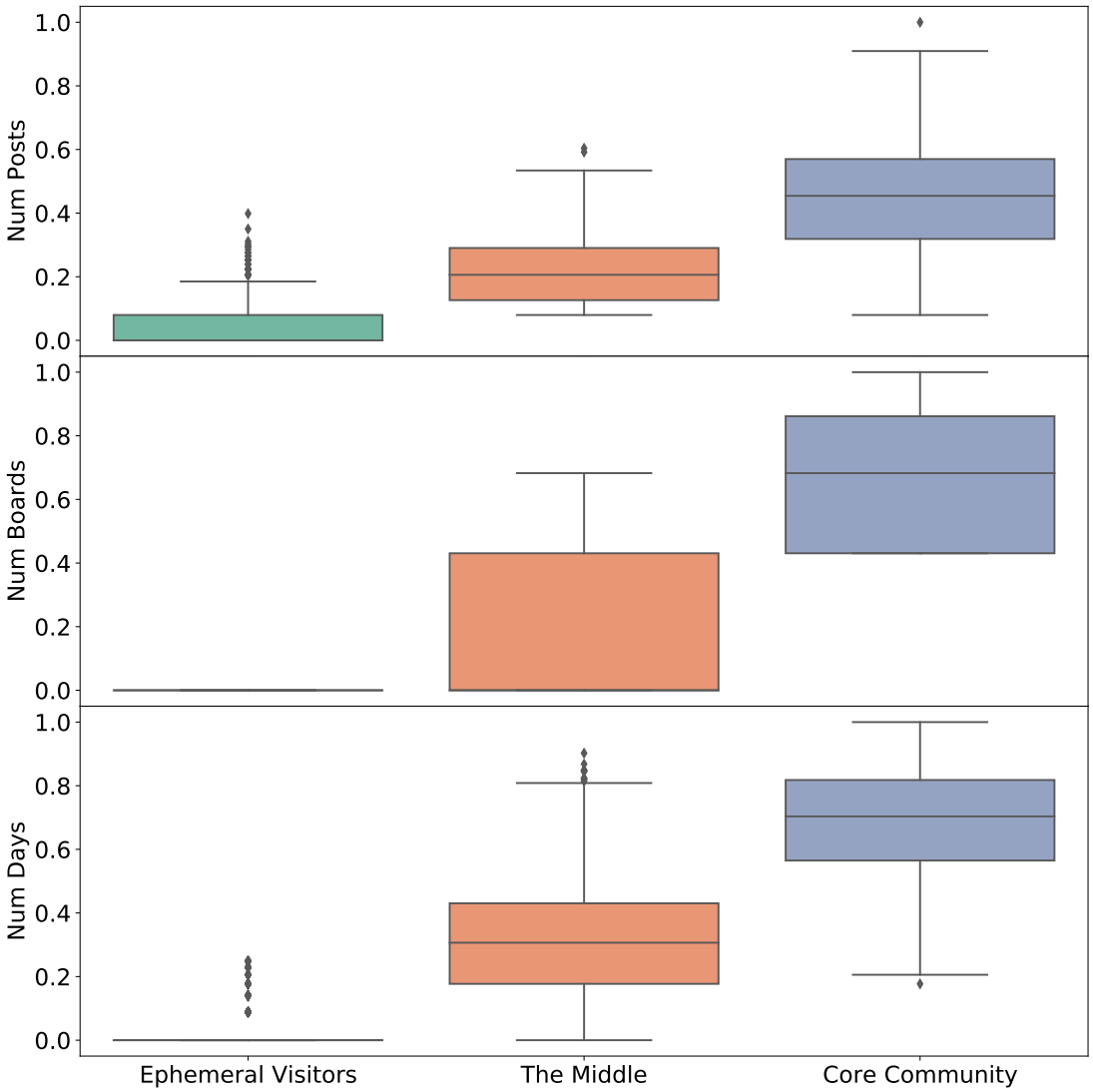


Figure 7.11: Box plots showing the distribution of each meta-feature across the three clusters.

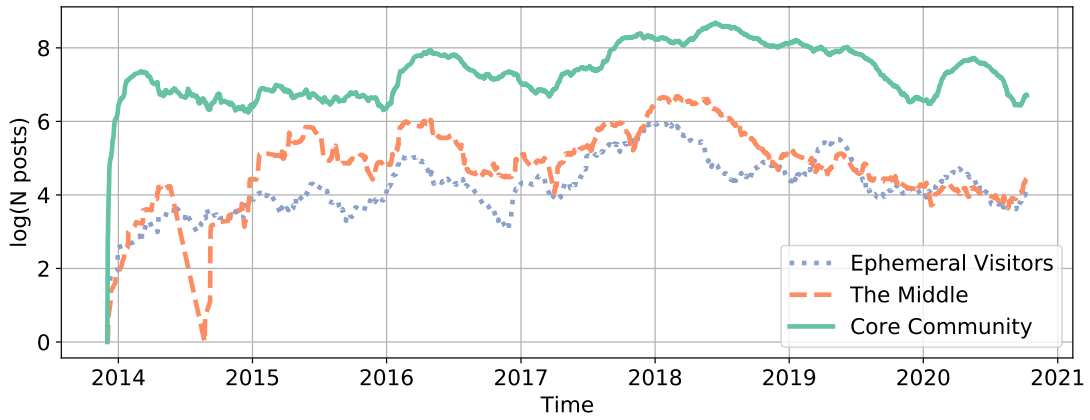


Figure 7.12: Plot showing the logged rolling frequency of posts for each meta-group, looking at 90 day windows.

The **Core Community** generally contributed more posts, across multiple boards, and spent more days on the forum. The median value of every feature is considerably higher than in the other two groups. This points to users in the Core Community being much more active and posting in a greater variety of boards, including the off-topic ones, suggesting that they engage in the more social aspects of the forum.

Looking at Figure 7.12, it seems that the posting behaviour over time is similar for all groups. There is not a notable surge of posts by ephemeral posters that is not also reflected in an increase in posts by all members. From this plot it is difficult to tell if any group initiated these increases, causing other groups to respond.

Based on this exploration, it would seem that the Ephemeral Visitor and Core Community groups are of most interest for comparison. The Middle is less well defined in its own right, and may be a place for users who do not fit into the other two groups.

7.3.2 Comparing the Language of the Groups

With the meta-groups defined, the next step is to investigate the differences in language usage between the different groups. Initially, this will be done by looking at the keywords of each group compared to every other group. The method for finding keywords was the same used in Section 6.6.1, where it is described in detail. By performing this analysis, we can better understand what role these groups play on the forum, and help explain the dynamics between the community, and visitors in this community.

Figure 7.13 shows the keywords of each group, with respect to each other group. Reading across the first row, for example, one can see word clouds[3] representing the words overused by users from the Ephemeral Visitor compared to the other groups. The size of the word in a word cloud corresponds to the Log-Ratio of the word, meaning that the larger it is, the more it is overused.

Starting with the keywords of Ephemeral Visitors to the Core Community (top right), it is immediately clear that many of the posts by members of this group involve introducing themselves. This is evidenced by words such as “hello” and “hi” being heavily overused. This is unsurprising, as introducing oneself is something one is likely

[3]For generating the word clouds, the *wordcloud* python package was used. (`https://pypi.org/project/wordcloud/`), as in Chapter 4

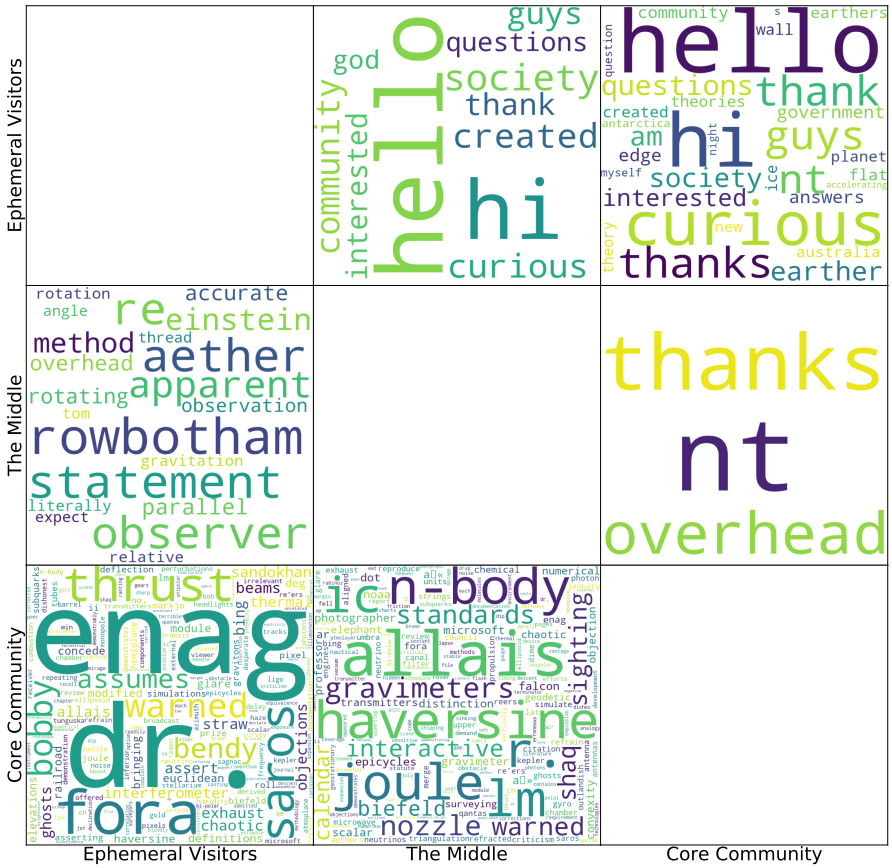


Figure 7.13: Keywords of each group, compared to each other group, in the Flat Earth Society Forum. Keywords classed as words with Log-Ratio >1 and frequency >100.

to do in one's first post on a forum, and these users mostly only post once, meaning that the majority of posts in this group will be introductions. Ephemeral Visitors also seem to be predominantly asking questions. "Interested", "curious", "thanks", and "answers" are all overused by this group. This gives an impression of users who post a question on the forum on a one-off basis.

Another observation is that certain well known Flat Earth concepts are overused by Ephemeral users compared to the core community. Foremost, both "flat" and "earthers" are overused by this group. This might suggest that the authors of these posts do not consider themselves flat earthers, though whether they are willing to be persuaded remains another question. The word "lie" being key may suggest that plenty are not. Widely publicised elements of Flat-Earth Theory are also overused by this group. "Antarctica" and "wall" are both key. These relate to the theory that the Earth's disc is surrounded by an ice wall, commonly known as Antarctica. "Australia" is also key. This may relate to a theory that was circulated widely online that Australia was not real, and had been invented as part of the round-earth conspiracy. All this may imply that visitors to the forum read about the most outrageous flat-earth theories online, and then come to the forum to ask questions.

The next keywords we will look at are those overused by the Core Community compared to Ephemeral Visitors (Bottom Left). Some of the keywords seem to relate to forum moderation. "Warned" and "fora" are undoubtedly key because they are frequently said by moderators and admins, who also happen to be regular posters, and therefore members of the Core Community. Many of the words that are highlighted here are technical terms or acronyms. Words such as "ENAG" (Earth Not a Globe), "interferometer", and "Saros" (the Saros cycle) all appear as key. This suggests that the Core community group use much more technical language than the Ephemeral Visitors. This makes a lot of sense given they are, having posted more often, more immersed in the world of Flat Earth Theory.

If core community users are defined by their use of technical language, and ephemeral visitors by their asking of questions, what can we learn by looking at the middle users? This group overuses technical terms compared to ephemeral users (e.g. "Rowbotham", "aether"), which suggests that it might include users who are part of the community. However, the core community still overuses more specific technical terms

(e.g. "Allais", "Haversine") compared to the middle users. This could suggest that these middle users use more general technical terms. They may be engaging with the scientific discussion, while not being aware of some of the niche concepts. The middle group also overuses the word "thanks" compared to the Core Community, which could suggest that the group contains users coming to ask questions, similar to the ephemeral group.

Our findings about the middle group suggest that it is made up of users who are less active on the forum, but still engage with Flat Earth debate, and slightly more active visitors who may be outside the community. It would be interesting to explore the beliefs of this set of users, to see if they are less-enthusiastic flat-earthers, or more-prolific round earthers. It is worth noting that these suggestions are based only on keywords for now, so they come with caveats. However, this is still interesting evidence that helps us further characterise this community.

7.3.3 Comparing the Language Models of Groups Over Time

So far we have looked at some basic linguistic differences between the groups we have identified. Now we will look at how the language models of the Core and Ephemeral groups change relative to themselves and each other. We excluded the Middle group, to create a more distinct separation between the groups, as the Middle group seems to contain users who do not quite belong in either of the other two. To do this comparison, we will use the ACE method, introduced in Chapter 5.

Figure 7.14 shows the unpredictability of each group's language model over time. This is a measure of the average cross-entropy between random samples of users from the group with the rest of the group, over 10 runs. The core community is consistently the least unpredictable, while the Ephemeral group fluctuates more in its unpredictability. The Ephemeral group has peaks of unpredictability in 2016 and 2018. These peaks correspond to rises in posting activity shown in Figure 6.6, suggesting that the language models of the group becomes less stable when there is an influx of new users. The core community becomes less unpredictable after the first of these peaks, but increases its unpredictability after the second, larger peak. This could suggest that the first peak stabilised language on the forum slightly, but when even more new users joined it became more unpredictable.

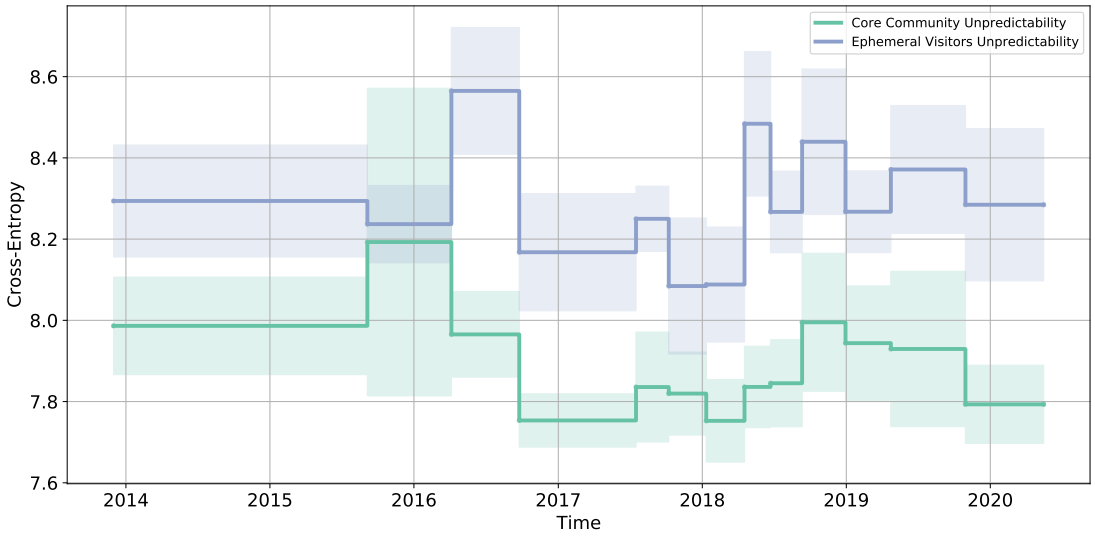


Figure 7.14: Unpredictability of each group. Window size 15000, window step 15000, 10 runs, not balanced, with no contribution limit. Posts split into chunks of 30 characters.

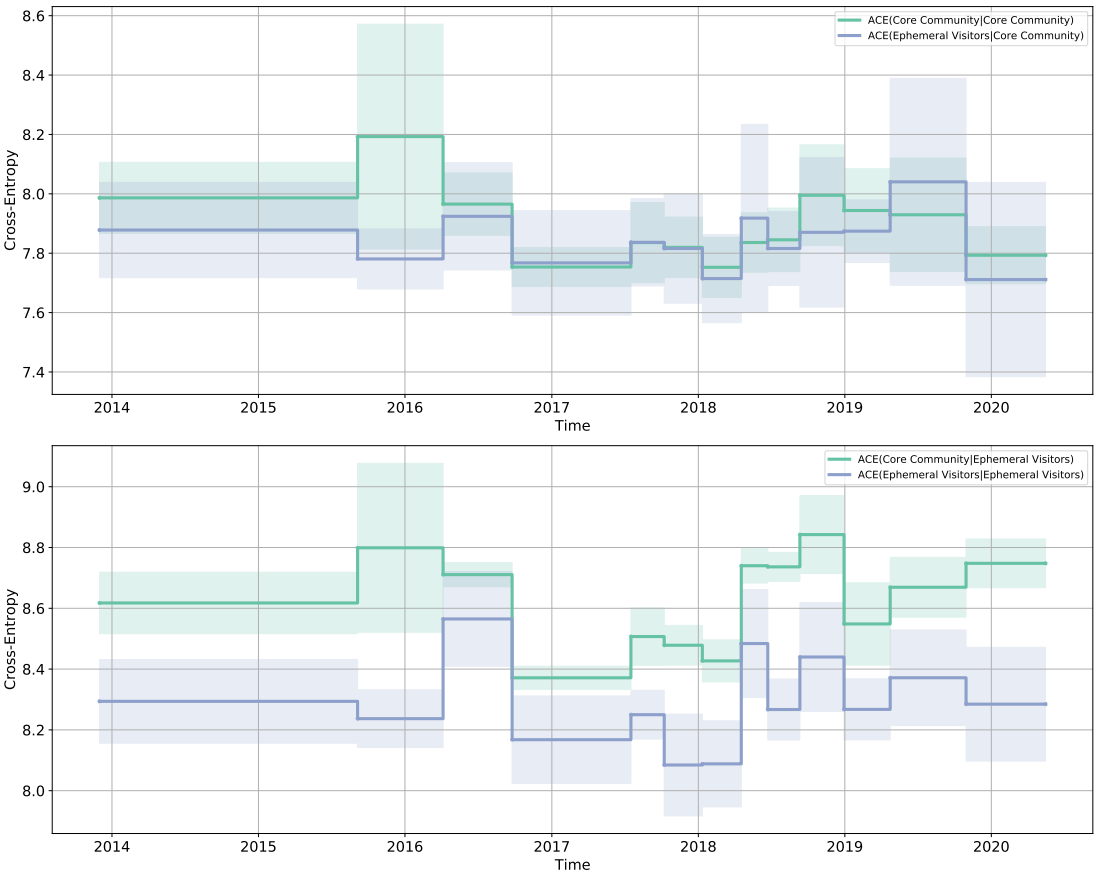


Figure 7.15: ACE of each group according to the snapshot model of each group. Window size 15,000, window step 15,000, 10 runs, not balanced, with no contribution limit. Posts split into chunks of 30 characters.

Figure 7.15 shows the ACE of each group to the snapshot of the other, compared to the unpredictability of the other group. The top graph does not tell us much, with both groups staying fairly similar over time. We certainly do not see any suggestion of convergence or divergence over time, so it is difficult to draw any conclusions from this about influence.

The lower graph does show a couple of interesting things. The core community becomes less divergent from the core group after the first peak of ephemeral unpredictability. It then has an increase in divergence at the same time as the second ephemeral peak. These changes appear to be more driven by the changing unpredictability of the ephemeral group, than by changes in the core group. This highlights a problem with ACE, and other entropy-based techniques, which is that it is difficult to tell whether the divergence is cause by a genuine change, or by the snapshot model becoming less predictable. One advantage of ACE, is that it at least highlights this problem by allowing one to easily show the unpredictability alongside the divergence.

The analysis has not provided much indication of influence or convergence at this level, though it has shown that the unpredictability of the groups fluctuates with changes in membership. A lack of influence would not necessarily be surprising given that the Ephemeral group in particular consists of completely different people at every time step. It is also possible that ACE may be more useful when consistently looking at the same users over a longer time period.

7.3.4 Concluding Thoughts on Meta-Groups

In this section, we have used three basic meta features to find potential groups on the forum. Of the three groups we identified, two seem to represent logical groupings: members of the core community, and ephemeral users who only spend a brief amount of time on the forum. The third group lies in between, showing there is no discrete cut-off between core and ephemeral members.

The range of analyses we could perform on these groups was limited by the huge imbalance in data quantity between them. Ephemeral users inherently contribute far less to the forum than core community members, so it is difficult to get much data for comparison. A more interesting question may be to see if groups exist within the core

community. Such groups may help to identify trolls, etc, on the forum. This would be difficult to do with posting based features, however, so we would need to find a way to cluster groups of users based on their language, as we will discuss in the next section.

7.4 Searching for Linguistic Groups

In Section 7.3, we looked broadly at different types of user on the forum by identifying groups of users based on their posting behaviour. We found that most of the users on the forum contributed very little, for a very brief window of time, while most of the posts were made by a smaller, core community of posters. While we did look at the keywords of these meta-groups, this did not tell us much about what separates different users, partly because ephemeral users contributed comparatively little text.

In this section, we aim to study the language of the core community and see if we can find types of user on the forum, based on their linguistic features. Looking more in depth at the features that separate frequent posters on the forum will help us answer RQ3 from Section 1.3. The groups we identify may correspond to interesting ideas such as belief in FET, or inform us about the style of discussion on the forum.

We chose only to look at the core community to better understand the differences between active members of the forum. The ephemeral users contribute so little individually, often a single post, that we cannot hope to learn much about different individuals by looking at their language. As a group, their linguistic features have already been discussed in Section 6.6. For this reason, we only considered the core community.

Because there are no group labels for the users on the forum, we used an unsupervised approach. Agglomerative hierarchical clustering was chosen because it is widely used, and allows the comparison of feature vectors with cosine similarity, a preferable distance measure for comparisons of language data to euclidean distance, the measure used by K-Means. Two feature sets were used: bag of words (BoW) and bag of part-of-speech trigrams (PoS-tri). BoW will tell us more about the surface level topical differences, while PoS-tri will suggest more stylistic differences.

For BoW, we looked at the 10,000 most common tokens in the FE sections of the forum, and for PoS-tri we looked at the most common 1000. We used TF-IDF values for

BoW, and counts normalised by document length for PoS-tri. As we have before in this thesis, we produced feature matrices for both these feature sets, and then standardised the data by subtracting the mean and scaling to unit variance. We then reduced the dimensionality of the data for clustering using PCA, which we used to reduce the features while preserving 95% of the variance. This resulted in 158 features for PoS-tri, and 261 for BoW. Only posts on FE boards were considered because we wanted to constrain the topic to FET as much as possible.

Initial clustering of the data produced fairly weak clusters, with a cophenetic correlation coefficient of 0.27 for PoS-Tri, and 0.31 for BoW [4]. The groupings may not be clear, but looking at them may still reveal interesting characteristics of users in the forum. For both feature sets we considered two clusters, as this number of clusters resulted in the highest silhouette score[5]. The clusters produced by the two feature sets were distinct, so we will need to look at both to understand how users are being split.

We will now look at the clusters formed using PoS Trigrams. In particular, we will look at the PoS trigrams and words overused by these groups. This was done using Log-Ratio, as in previous sections. The key PoS Trigrams suggest a split between technical and casual discussion. Punctuation and numbers dominate for one cluster, while a more casual language style (full of interjections and pronouns) defines the other. This suggests that there are some users who engage in more personal dialogue on the forum, and those who write more scientifically/formally.

The keywords of the PoS-tri clusters also demonstrate this behaviour, with the technical cluster containing more technical words (e.g. "longitudinal", "meridian", "ether"), and the casual cluster being made up of more argumentative language, overusing words such as "angry", "strawman", and "shit". Interestingly, the casual cluster also overuses several terms such as "ban", "warning", and "fora", that suggest it contains moderators on the forum. In fact, the casual cluster does contain more position-holding members of the forum, which may mean that this cluster is comprised of arguments between visitors and moderators. It is interesting that the moderators are not necessarily the users laying out complex mathematical arguments, etc. Technical discussion is seemingly not limited to those who organise the forum. The two clusters appear to represent two different sides to the forum, one side comprising arguments

[4] "Good" clusters would be close to 1.

[5] Though the score was still close to zero for both feature sets, indicating overlapping clusters.

between users, and another containing more detailed discussion.

Next we looked at the clusters produced by the BoW features. These clusters seemed to correspond much more clearly to a user's prevalence on the forum, with the higher posting users tending to be clustered together. This could suggest that the top users in the community are more linked by their topical, rather than stylistic, features.

The keywords of these clusters display similar behaviour to the PoS-tri clusters. One cluster overuses more technical terms (e.g. "molecules", "gravimeter", "measurable"), which corresponds to the technical PoS-tri cluster. This is the cluster that contains more prominent users. Another interesting keyword for this cluster is "scripture". which is mixed in with more scientific sounding words. This could potentially suggest that this group contains the FE believers on the forum, as previous work has found that FET was associated with religious beliefs [Mohammed, 2019, Landrum and Olshansky, 2020].

The other cluster contains keywords that suggest much less detailed discussion of FET. Similar to what we found looking at Ephemeral users, this group discussed the "Ice Wall" more than the other cluster. This could suggest that they are outsiders coming in, having read about some of the more outlandish aspects of FET online. These users also overused words such as "prove", "believe", and "fake", which suggest they may be round Earthers visiting the forum to question, or attempt to disprove, the beliefs of the community.

7.5 A Closer Look at the Top 20 Users

So far, this analysis has provided some interesting insights into the community. We have observed a split between technical and casual discussion, and another split that seems to separate users coming to question FE believers from the more active members of the community. In this section, we will perform a more granular analysis of the 20 highest posting users. This is an interesting group because they contribute a large proportion of the forum's posts, accounting for 42% of activity on the FE boards. Looking more closely at this group will tell us about the language used by the most prolific members of the community. All these users were clustered together by BoW in Section 7.4, in the cluster that contained more technical FE discussion. Conveniently, this group is also manageable enough to manually label the FE belief of users, which will allow us to gain

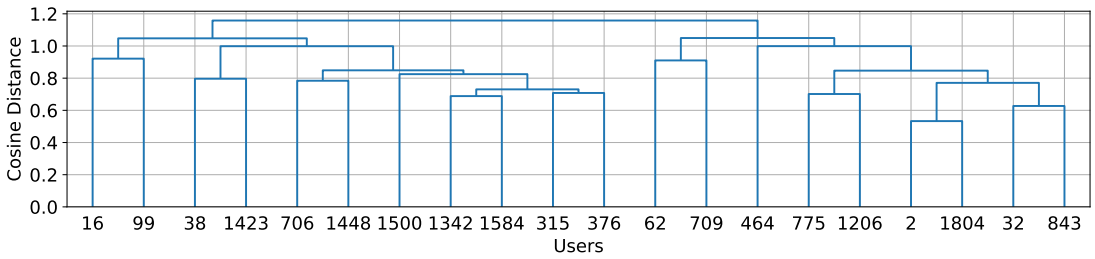


Figure 7.16: Dendrogram showing the clusters of users made in hierarchical clustering on the PoS-trigram feature set. For linkage we used average, and cosine distance was used as the metric.

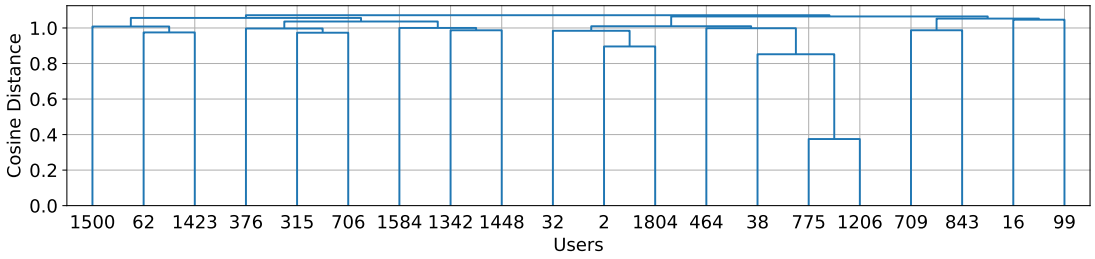


Figure 7.17: Dendrogram showing the clusters of users made in hierarchical clustering on the TF-IDF BoW feature set. For linkage we used average, and cosine distance was used as the metric.

new insight into our data.

As in Section 7.4, we used hierarchical clustering using two feature sets (Pos-Tri and BoW) to split the users into clusters. Figures 7.16 and 7.17 show the dendrograms produced by this clustering. Immediately we can see that the clusters are not particularly well defined, most users only clustering at fairly high cosine distances. The Pos-Tri features cluster better than BoW, but even so the cophenetic correlation coefficient is 0.47. The clusters are better defined than they were for the entire core community. Again, we considered two clusters. The assigned clusters for each feature set are shown in Table 7.4.

Unlike the clusters for the entire core community from Section 7.4, the PoS-Tri and BoW clusters are largely the same (for 16 out of the 20), suggesting that similar differences are being picked up by both style and content features. The key PoS-Trigrams of the PoS-Tri clusters show the same phenomenon as for the entire core community, splitting users engaging in technical discussion from those using more casual language. Keywords are also similar, with one cluster overusing words relating to theoretical concepts (e.g. “astronomical”, “longitude”), while the other overuses terms relating to moderation (e.g. “fora”, “warned”) and more argumentative language (e.g.

Table 7.4: Table showing the clusters for each of the top 20 users, according to hierarchical clustering performed with two feature sets, as well as labels for Flat Earth belief, and whether or not the user holds a position on the forum. Users are ordered from top to bottom by number of FE posts.

PoS-Tri	BoW	Belief	Position
1	0	FE	True
0	0	FE	True
0	0	RE	False
1	1	RE	False
0	0	FE	True
0	0	UNK	False
1	1	RE	False
0	0	RE	False
1	1	RE	False
1	0	RE	True
1	1	RE	False
0	1	RE	False
1	1	UNK	False
1	1	RE	False
1	1	FE	False
0	0	RE	False
0	0	RE	False
1	1	RE	False
0	0	RE	False
1	0	RE	False

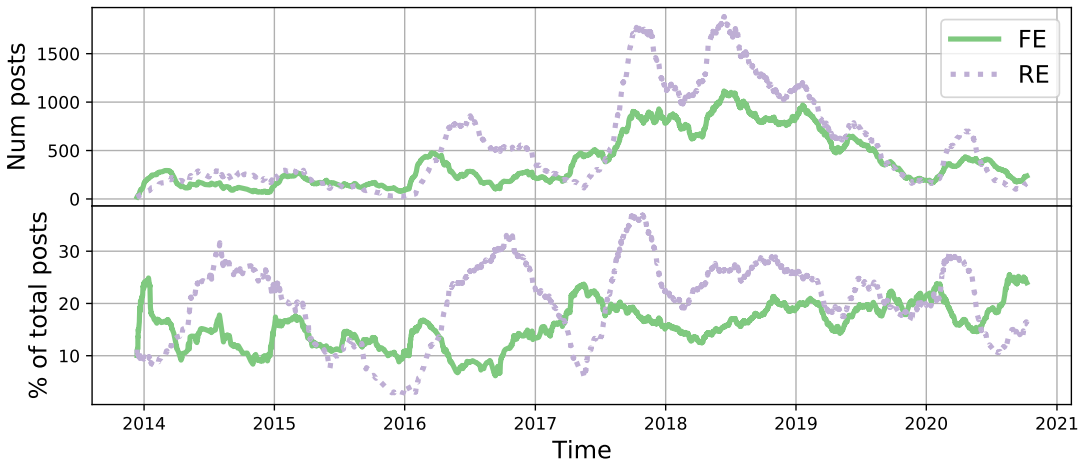


Figure 7.18: Plot showing the number of posts for FE and RE users (in the top 20) over time, using a 90 day rolling window. Top plot shows raw number of posts, and bottom shows it as a percentage of all FE posts.

"silly", "nonsense").

Table 7.4 also shows which of the users in the top 20 hold positions on the forum. Within this group, only four of the 20 hold positions. These four only make up a third of the total number of members who hold positions on the forum. It is interesting that a larger proportion of these users are not found in the top 20, as it suggests the most active members are not all involved in the organisation of the community.

Even though one of the PoS-Tri clusters overuses moderation-related terms, the position-holding users are evenly split between them. For the BoW clusters, the casual language cluster contains all of the position-holding users. This may suggest that position-holding users are more defined in their language by content features, as opposed to style. It is interesting that the position-holding users are not associated with the technical language cluster, given that one would expect these users to be the most devoted Flat Earthers in the community.

Relation to Flat Earth Belief

To ascertain the belief of users, we sampled 50 posts from each of the twenty users, and manually read through them to identify their stance on the Flat Earth. Users were labelled as FE (Flat Earth), RE (Round Earth), or UNK (Unknown) based on their stated positions. Where it was unclear, we categorised a user as UNK. These labels are far from perfect, particularly as one's stated belief does not necessarily reflect reality, but they

will give us some indication of belief. Table 7.4 shows the belief of each user alongside their clusters and whether or not they hold a position.

Interestingly, out of the top 20 users, only 4 were clearly FE believers. While we were fully aware that the forum was a place for discussion from both sides, it is interesting that FE believers are so outnumbered in the group we are looking at. Though a smaller group, the FE users generally contribute more posts.

The groups of FE and RE and users we identified do not align neatly with the clusters. For the PoS clusters, two FE users are found in each cluster. This suggests it is not only FE users who engage in complex discussions, and not all RE users come to the forum to troll and mock. It is interesting that both RE and FE users engage in both casual and technical discussion, because it suggests that the community is more complex than simply a small group of FE believers skirmishing endlessly with ephemeral RE'ers. For the BoW clusters, the three most prolific FE users, all of whom hold positions on the forum, are in the same cluster. This may suggest that content features differentiate the belief of users more than style features. These results indicate that the style of discussion is not inherently different between flat and round Earth believers.

One surprising observation is that one of the four users with positions is not a flat Earther. This is potentially very interesting, as it could show that RE users are at the very core of the community. On closer inspection of the data, this particular user's position was listed as "Purgatory". It is unclear what this means, but it could mean that, though they have a position on the forum and have been active for most of the forum's lifetime, they are not a moderator.

Figure 7.18 shows the number of posts for FE and RE users over time, which shows how the RE users have peaks in posting that correspond with the peaks of new users shown in Figure 6.9. Figures 7.19 and 7.20 break this down for individual users in the top 20. It shows that FE users have been more active over a longer period of time. Only one FE user joined after the forum's creation. RE users, on the other hand, were mostly active for shorter periods of time. Most joined from 2016 onwards, the period in which many new users came to the forum, as shown in Figure 6.9. Even so, many of these RE users stayed for over a year, not an insignificant amount of time. These findings back up the idea that while originally the forum was largely comprised of FE users, as the theory became mainstream, more and more round Earthers visited the site. The length of some

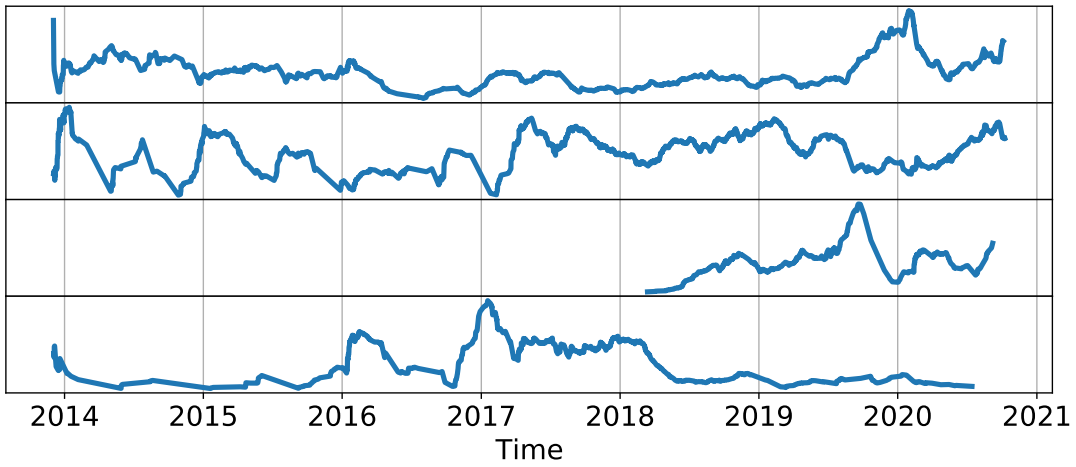


Figure 7.19: Figure showing the posts over time for each of the FE users in the top 20.

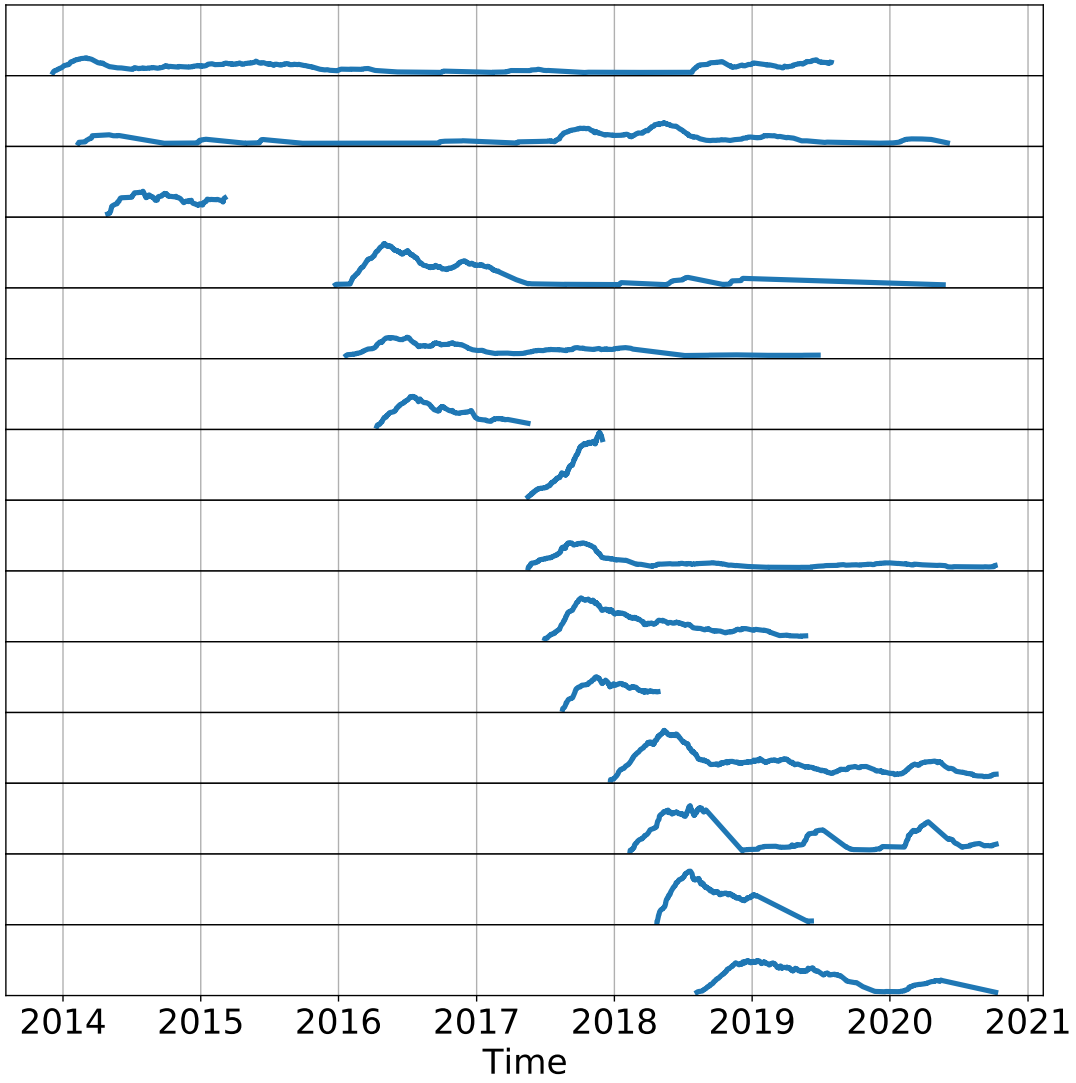


Figure 7.20: Figure showing the posts over time for each of the RE users in the top 20.

of these RE users' stays suggest, however, that their relationship with the community was not limited to a fleeting curiosity, or trolling.

We also looked at the key words of the FE and RE users we identified. Terms relating to moderation were overused by the FE group, suggesting that, unsurprisingly, the moderators purport to believe the Earth is flat. There no longer remains a split between use of technical terms; some technical words are overused by both groups. Words such as "tide" and "sunsets", subjects one might bring up if trying to prove the Earth's sphericity, were overused by the RE group. These findings, along with the key features of our clusters, suggests that the style of debate does not seem to indicate belief in the flat Earth, and that users of differing beliefs engage in both technical and non-technical discussion on the forum.

7.6 Discussion

In this chapter, we have expanded on the work in Chapter 6 by performing more detailed analyses on the Flat Earth forum. Specifically we have looked at the way language changes in this community, and identified groups of users based on meta-information and linguistic features. This has helped us to answer RQ2 and RQ3 from Section 1.3. In this section, we will summarise the key things we have learned throughout this chapter.

Language Change in Flat Earth Communities

In Section 7.2, we looked at language change over time within, and between, flat Earth communities. This provided an opportunity to test methods from Chapter 4, and contribute to answering RQ2. It also helped to further our understanding of how conspiracy communities change over time, helping to answer RQ3.

Looking at language change, we identified some words that may have changed in usage over time on the FES forum. For example, the word "ice" became more prevalent over time. Based on what we have discussed about ephemeral users favouring this discussion topic, it is possible that this suggests the community gradually included more of this type of discussion. Despite this, most of the words we looked at were stable in their usage, suggesting that the central ideas of Flat Earth theory have been fairly constant during the time period we looked at. We also observed some terms and phrases

developing. For example, the phrase "ice wall" became dominant in later stages of the forum's life, while early on there were alternatives such as "ice ring".

Using Variability-based Neighbour Clustering, we identified epochs based on peaks and troughs of popularity. This was useful for finding trends in the forum. For example, we could observe the word "Shaq" becoming popular, referring to a celebrity who declared themselves an FE believer. Around the same time, references to "FE'ers" became more common, possibly suggesting an influx of new Round Earther users.

We witnessed some other possible trends in the forum. Certain keywords, e.g. "aether" became much less popular over time, possibly showing the way concepts evolve. The forum's hierarchy, particularly the council running it, was much more prevalent towards the start of the forum's life. This analysis has highlighted some interesting potential routes for future investigation.

Finally, we compared the FES forum to FE subreddits using cross-entropy based methods. We found the community was more divergent from some Flat-Earth communities than others. Divergence often occurred with an influx or decrease in popularity, though it was unclear if this was caused by a change in data quantity, or in the language of the communities.

We also compared the forum to related communities `r/science` and `r/conspiracy`, and found that the forum was less divergent from `r/science`. Though the effect was small, this could suggest that the language of conspiracy is not universal. The forum being more similar to scientific discourse lines up with our findings that FE debate involves complex, technical language.

Finding Meta-Groups

Initially we looked for groups in the forum using posting information and meta-data. By doing this we learned some interesting things about posting behaviour on the forum, and contributed to answering RQ3 by learning more about the behaviour of conspiracy communities. For example, most users on the forum post only a handful of times, usually over the course of a single day. At the same time, a small active group of users make up the bulk of the forum's contributions. These users are the only ones who use the forum as a more general social space, posting in the off-topic sections. This is clearly a community, not simply a place where people briefly come to mock Flat Earthers, though

the high turnover of new users suggests it may also be that.

We identified two main groupings of users: ephemeral visitors, and the core community. There were also many users floating between these groups. These users may be ripe for further study.

A keyword analysis was performed on these groupings. We found that ephemeral users asked a lot of questions, and referred to FE believers in ways that seemed to imply they were not one themselves: e.g. "FE'ers". This suggests that many of them were outsiders coming in, rather than people who already believed the Earth was flat. This group also mentioned well-publicised FE concepts such as the "ice wall" and "Australia" (or lack thereof[6]). This could suggest that these are users who have read about the FE community online and come along to ask questions or challenge their beliefs.

The core-community used many technical terms, as well as words relating to forum administration. This is what you would expect from the central group on a forum. Many of the users floating in between the two groups also engaged in some technical discussion, though lacked some of the particularly niche vocabulary.

We looked at the relative linguistic change of these groups over time, but found limited evidence of influence. This could be because of the high user turnover; there is simply not enough time for new users to be influenced before they leave. While we did not see much influence, ACE did show how the language of groups had spikes of unpredictability during periods when many new users joined the forum. Our analysis provided a good opportunity to test the ACE method from Chapter 5 on a new dataset, contributing towards RQ2.

Finding Linguistic Groups Within the Core Community

Following on from this analysis, we also looked for groups based on linguistic features within the core community, to learn more about the types of user who are active on the forum. We did this with hierarchical clustering using BoW and PoS-trigram features.

The PoS-Trigram clusters suggested different styles of discussion within the group: those engaging in technical/formal discussion, and those writing more casually. Users in the technical cluster used more scientific vocabulary, while those in the casual cluster

[6] https://www.theguardian.com/technology/shortcuts/2018/apr/15/australia-doesnt-exist-and-other-bizarre-geographic-conspiracies-that-wont-go-away

used argumentative language. The casual cluster also contained moderators from the forum, suggesting that it may be made up of RE users arguing with the site's moderators.

The BoW clusters suggested a difference between less and more prominent users. More prolific users engaged in more complex discussion, while less prolific users appeared to use the forum to question and challenge flat Earth theory.

The results from this clustering contributed to answering our research questions in several ways. We contributed to RQ2 by showing how unsupervised clustering methods can be useful for breaking down the language of communities. In terms of RQ3, the findings of this analysis painted a picture of the FE community as a place where some users debate theory, while others argue. We also addressed RQ1, showing that discussion of false information can comprise multiple different styles of language.

A Closer Look at the Top-20

In Section 7.5, we looked in more detail at the top 20 users on the forum, clustering them to find the linguistic features separating these highly active users. As in Section 7.4, the users were split between casual and technical language. We also found that the position-holding members of the forum were more associated with the casual cluster, suggesting that it is not the forum's organisers who predominantly engage in technical discussion.

Based on our comparison of these user clusters to manually assigned labels of belief, it did not seem that the clusters corresponded to Flat Earth belief. Both FE and RE users engaged in technical discussion, and RE users were not separated from moderators. Interestingly, RE users outnumbered FE believers in the top 20, but FE users were more consistent posters over time. Our findings suggest that the features of FE and RE belief are less distinguishing amongst the prolific posters on the forum than the style of discussion they engage in.

These results reveal that the Flat Earth community is not as simple as one might think. Not all of the active community members believe that the Earth is flat, and many of the RE believing members engage in technical discussion, no less than the FE members. It is also surprising that there were more RE believers than FE in the top 20 users. The community clearly does not only serve as a space for FE believers, and the round Earthers present are not all there simply to troll or abuse.

In so far as answering RQ3 goes, we have shown that false information communities

can be complex communities, that do not only serve to spread false information, but also serve as more general social spaces. Our findings also suggest that the users of these communities cannot be taken for granted as believers in whatever false information the community is based around. Many non-believers participate in discussion, and not all of them seem to do so as trolls or abusers.

The linguistic analyses we performed also have implications for RQ1. We found that the style of discussion was a better separator between the top users on the forum than their belief was. Both believers and non-believers engaged in similar styles of discussion. This shows that analysing communities such as this does not sit as comfortably in the study of false information, as other media such as fake news. It also highlights how important it is that future works in false information think about belief, and avoid the assumption that entire communities of people have the same beliefs.

7.7 Conclusion

In this chapter we have performed a deeper linguistic analysis of the Flat Earth communities introduced in Chapter 6. We performed a diachronic analysis, using methods from Chapter 4, looking at the forum in the context of other Flat Earth and related communities. Posting patterns were observed over time, with clear peaks possibly related to mainstream interest in Flat Earth theory. We identified linguistic stages in these communities, and identified words that changed in meaning over time. Finally, we compared various FE and related communities, and found that certain communities diverged more than others from the language model of the FES forum. Future work will be needed to establish the underlying reasons behind these differences.

We have used meta and linguistic features to identify meaningful groups of users within the community. This analysis showed that the forum is split between ephemeral users who spend very little time on the forum, and the core community of users who contribute more substantially. We dived deeper into potential linguistically defined groups within the core community, and found that some users were more technical in their discussion, while others were more casual. The groups did not, however, correspond to FE belief, suggesting that users are more defined by their style of discussion than their belief.

Our findings have painted a more complex picture than one might expect of the FE community. The FE forum does not simply serve as a place for trolls to deride the beliefs of the community, though that behaviour does exist. Both believers and non-believers in the conspiracy theory make up the core community, and both groups engage in both technical and casual discussion. What we have learned in this chapter has provided a firm basis for future research into Flat Earth communities, and can be used more generally to better understand disinformation and conspiracy theories.

Chapter 8

Conclusion

This chapter will wrap up the threads of discussion explored in the thesis. It will begin by summarising each chapter in turn. The research questions from Chapter 1 will then be revisited, and we will assess the extent to which each was answered. We will then discuss the key contributions made in this thesis. Finally, possible future directions of research will be described, which could be pursued to expand on this work.

8.1 Summary

Overall the focus of the thesis was on creating new datasets and methods for the analysis of false information. In Chapter 2, we performed a survey of research from NLP and corpus linguistics relating to false information, language change, and online communities. This fusion of topics provided a wide context for the work described in this thesis. The review guided our choice of approaches to the problems we faced.

Chapter 3 described a case study of false information, introducing a novel dataset of April Fools (AF) hoaxes. We performed a feature-driven analysis of AF news articles, to better understand the features of false information. Classifiers were trained using a range of features to distinguish AF from genuine texts. The results of these classifiers, as well as their feature weights, were used to identify key features. We then identified features that were generalisable from April Fools to "Fake News".

In Chapter 4, we created a toolbox of methods for looking at language change over short time-spans within communities. This involved surveying a selection of language change methods, and proposing adaptations for the comparison of sub-groups over time.

The methods were tested using a dataset of UK parliamentary debates, to demonstrate their suitability for comparing the language of groups over short time-spans. This toolbox has applications for researching many kinds of community, including those dedicated to false information.

Chapter 5 then introduced a novel method called Average Cross-Entropy (ACE), designed specifically for comparing the relative language change of groups to one another over time. The usefulness of ACE was assessed by comparing groups of MPs over the course of the Brexit process. As with the toolbox described in the previous chapter, ACE was created with the analysis of false information communities in mind.

Chapters 6 and 7 brought together what we had learned from April Fools, and the toolbox of methods we had developed for looking at language change, and applied them to an online conspiracy community. Chapter 6 described the creation of a novel dataset of online Flat Earth communities. A meta analysis was performed to better understand and measure the shape of these communities. Using the Flat Earth Society forum, we then performed a linguistic analysis to characterise the language of Flat Earth debate, and how it differed from general discussion within the forum.

Finally, Chapter 7 described several more complex analyses of the dataset, relating to language change and the discovery of sub-groups. Methods from Chapter 4 were applied to the FE communities to see how they changed over time, both individually and relative to one another. We also looked for sub-groups within the community based on meta-information and linguistic features. We identified linguistic features that characterised the groups as a further way of learning about the types of user resident on the forum. These groups were also discussed in relation to Flat Earth belief, and future research directions were identified.

8.2 Research Questions

We will now revisit the research questions introduced in Chapter 1, to discuss the ways in which they have been answered.

RQ1: How can we increase our understanding of the language of false information by looking at previously unstudied sources?

In Chapter 3 we found some answers to this question. We found that features relating to detail and complexity were important in classifying April Fools hoaxes, a form of false information. These features also differentiated "fake" news from genuine, suggesting that looking at structural complexity of texts is important for identifying false information more generally. The results were promising, and indicated that some of the features we looked at were generalisable to multiple forms of false information. We also performed a corpus linguistic analysis, looking for key words and parts-of-speech which characterised April Fools and fake news. This analysis supported our findings, suggesting that genuine articles established more detail and false stories struck a more casual tone. Real news articles being more complex and detailed, and hoaxes using more vague language, fits with the idea that deception is less complex because of the increased cognitive load of lying [Carlson et al., 2004].

The findings from Chapter 6 provide an interesting comparison, with Flat Earth debate containing complex, technical language. There are many possible reasons for this. It could align with the findings of Markowitz and Hancock [2014], which showed that fraudulent texts overused genre-specific vocabulary, or it could be because false information communities cover a wide range of individuals and styles of discussion, with debate between believers, trolls, and sceptics. Another finding from Chapter 6 relating to RQ1 was that removing function words removed important terms from the corpus, such as the word "round". Our findings show that the language of false information is complex, and we need to consider the make up of the communities in which it is discussed, as well as the topic. While there are similarities between different types of false information, they cannot necessarily be relied upon in every circumstance.

Chapter 7 found that both believers and non-believers in FET participated on the forum, and used similar styles of discussion. This highlights that the language of believers in conspiracy theories may not always be different from that of non-believers, which shows the need to consider the true belief of individuals when building tools to detect false information.

RQ2: What methods allow us to observe the language of groups within communities, particularly regarding language change over short time-spans?

Chapter 2 surveyed various current approaches to community analysis. Based on this, Chapter 4 identified several relevant methods for observing language change, and adapted them for comparing sub-groups within communities over short time-spans, producing a toolbox of methods. In their original usage, most of these methods were applied to datasets spanning many years and were intended to look for changes in general language, rather than the language of specific authors or groups. We found that various existing methods could be adapted to compare the language of groups over relatively short time-spans. Each method served a slightly different purpose, and none individually was enough to gain a complete impression of the language change of a community. For example, VNC was useful for splitting a corpus into epochs, but was not useful for identifying particular events. The pros and cons of each method were discussed throughout the chapter.

Chapter 5 then introduced a new method, ACE, for comparing groups over time. This method built on an existing technique, Cross-Entropy, by using repeated sampling to avoid a group's language model being overwhelmed by individuals. It also allowed the estimation of the unpredictability of a group's language. The method was tested on a corpus of UK House of Commons debates during the Brexit period. Our findings were promising, but future work will be needed to test the method on new datasets and establish the boundaries of its usefulness. We used ACE in Chapter 7 to compare sub-groups within a Flat Earth forum, and though we did not observe much influence between these groups, we observed interesting changes based on unpredictability, and highlighted the generalisability of the technique in a new setting.

The methods introduced in Chapter 4 were also applied in Chapter 7 to look at language change in Flat Earth communities. We used diachronic embeddings and UFA to look at changing word usage in the community. Epochs were identified using VNC, that helped us understand the way that FE communities developed. We also demonstrated how cross-entropy could be used to plot the convergence and divergence of these communities over time, compared to each other, as well as non-FE subreddits. Though we found language to be fairly consistent, the methods still highlighted interesting behaviours that pave the way for future investigation.

RQ3: What features characterise false information communities, and how do they compare to other communities?

Chapter 6 looked at the various meta-features of online Flat Earth communities. We found that these communities were similar to non-FE communities in many ways, such as having a small group of dedicated users who contribute the most. FE communities did, however, seem to have a smaller proportion of long-term members than the non-FE groups. This suggests a higher number of users who do not stay active for a long period. These may be trolls or curious visitors who have read about the theory online.

On Reddit, various FE communities have come and gone over the past decade. The surges in popularity of these groups may lend clues to why people join these subreddits. In Section 7.2, our findings suggested that these new users left their mark on the language use of the forum, with more words appearing that suggested "Round Earthers" joining the community.

An interesting finding while looking at the Flat Earth Society (FES) forum was that it served not only as a place for FE debate, but also as a general social space. When looking at the volume of Flat Earth and off-topic posts over time, it became apparent that early in its life the forum was mostly general discussion, but as Flat Earth Theory came into popular awareness, Flat Earth discussion became dominant. This possibly fits with the narrative that FE communities have been taken over by Round Earthers.

Chapter 6 also involved an analysis of the linguistic features that characterise Flat Earth debate. This was done by comparing the FE sections of the forum to off-topic areas. We found that the community had developed its own technical vocabulary for discussing Flat Earth Theory. These detailed, technical arguments may appear indistinguishable from normal scientific discourse if context were removed. Previous work has associated conspiratorial belief with low levels of analytic thinking [Goertzel, 1994], despite the convoluted arguments put forward by believers.

Chapter 7 dived deeper into the linguistic features of the FES forum. This included a look at language change within the community, and between the community and a number of subreddits. While our findings suggested that there were no obvious large scale changes in the forum's language, we still saw some interesting behaviours. For example, evidence seemed to suggest that there might be some truth in the idea that Round Earthers became increasingly prominent over time. We also found that the `r/`

`science` subreddit was less divergent than `r/conspiracy` from the FES forum. This may suggest that there is no generic language of conspiracy, and supports our previous finding that language on this forum appears scientific on a surface level.

Section 7.3 investigated groups of users on the forum based on posting activity. Clustering suggested three types of user: ephemeral users, the core community, and users in the middle. Ephemeral users referenced certain well known FE concepts, while the core community used more technical language. The fact that ephemeral users overused terms relating to widely publicised elements of Flat Earth Theory suggested that they were users who had posted on the forum having read about FET online.

The linguistic analyses we could perform on these groups were somewhat limited by grossly imbalanced data: the ephemeral users contributed very little text compared to the core community despite substantially outnumbering them. So for Section 7.4 we looked only at the core community, looking for groups amongst them based on their linguistic features. This analysis revealed a split between users who made use of technical language, and those who wrote more casually. Keywords also suggested that even within the core community, there was a split between users engaging in complex discussion of FET, and more argumentative users.

We also performed an analysis of belief on a subset of the 20 most prolific users on the forum. This analysis showed that the majority were Round Earth advocates. It is interesting that so many of the most prominent users in a conspiracy community would not voice support for the community's driving theory. Even so, the FE users were active more consistently over the full span of the forum's existence, suggesting that Round Earth users participate in bursts, possibly while arguing about FET holds their interest.

When we clustered this group of users, they did not cluster based on belief, and remained to be identified by casual or detailed language. This suggests that the style of discussion engaged in has more effect on a user's language than their belief. Both FE and RE users engaged in both these styles, showing the fine line separating them.

Overall, our analysis paints a picture of a community made up largely of passing users, who briefly engage with the forum to ask questions to regular members or point out flaws in their belief. At the centre of the community there are a small group made up of both believers and sceptics, who debate Flat Earth ideas. Future work will need to go into more detail in studying the specifics of the discourse to shed more light on how

the community behaves, and compare it to other conspiracy communities to see if these behaviours are universal.

8.3 Contributions

Here, we will highlight all the main contributions of this thesis.

Created the first corpus of April Fools news articles

In Chapter 3, we introduced a new corpus for the study of disinformation. This corpus can be used as a verifiably, and completely, false set of news stories, where the intent of the authors is known, to compare to both fake and genuine news.

Performed the first NLP analysis of April Fools hoaxes

We also performed an analysis using this April Fools data. Our classification of hoaxes revealed some features that were useful in predicting both April Fools and fake news. Structural complexity was found as an important feature that distinguishes the two. Genuine articles also established more details than their inauthentic counterparts.

Built a toolbox of methods for looking at language change of groups within communities over short time-spans

In Chapter 4, we produced a toolbox of methods for observing the language change of groups within communities over short time ranges. We presented a novel survey of several language change methods from NLP and corpus linguistics, and discussed the suitability of each method for the comparison of groups over a short time-span. These techniques were used to reveal interesting insights into parliamentary debate, and the groups that comprise parliament.

Proposed a new method for comparing the language of groups over time

Chapter 5 introduced a new method: Average Cross Entropy (ACE). ACE is an adaptation of the well-used cross-entropy method for plotting language change. It adapts this technique to the task of comparing groups of users over short time periods. This is achieved by calculating cross-entropy over multiple runs with different samples

of users. We used a case-study of Brexit debates in UK parliament to demonstrate how this method could be used, and gained insight into the way different groups of MPs changed in their language over the Brexit period.

Created the first dataset of online Flat Earth discussion

Chapter 6 introduced a dataset of online Flat Earth communities. This is the first corpus of this particular topic, and one of the few corpora relating to online conspiracy theories. The data contained will be of use to researchers analysing disinformation, and online communities more generally.

Performed the first NLP analysis of Flat Earth debate and communities

As well as introducing the dataset, Chapter 6 also described the first NLP analysis into the language of Flat Earth debate. By looking for key terms on the FE sections of the FES forum, we showed some of the ways in which language is used in these communities. Chapter 7 then applied more complex analyses. Groups were compared to each other over time, as well as to external communities. Possible epochs were highlighted for these fora, along with the key terms that defined each period. We also explored sub-groups within the community, both using meta-information and linguistic features. The analyses contained within these chapters serves as a springboard for future research into online conspiracy communities and false information.

8.4 Future Work

To conclude, we will suggest some future directions for research that would build upon the work described in this thesis.

More datasets and new communities

This thesis has introduced three datasets in total, two of which relate to false information. The data has come from four different sources: news websites, parliamentary debate, online fora, and Reddit. We have shown how looking beyond common data sources such as Twitter can help to expand our understanding of false information. Future work should look for new sources of false information, and a wider range of

communities, which would help to test the generalisability of models trained to detect false information.

To better understand online conspiracy communities, it will also be important to analyse communities dedicated to a wide range of conspiracy theories. This thesis has covered one, the Flat Earth, but there are many popular conspiracy theories with substantial followings. Creating corpora for communities such as anti-vaxxers, climate change deniers, and QAnon will help to create a wider understanding of how these communities function, and what common features they share.

Labelling belief

Throughout the thesis, we mentioned our interest in belief, specifically in how the language of those who believe what they are saying differs from those who do not. This is a very difficult problem to tackle, as it is very hard to establish the true belief of an individual online. In Chapter 3, we got round this problem by choosing a dataset which was verifiably false, and where the belief of the author was clear. This was more difficult with the Flat Earth dataset, where we had no way to verify the belief of users, so we used unsupervised techniques to establish sub-groups.

At the end of Chapter 7, the 20 most prolific users on the forum were labelled with their projected belief. This supplemented the analysis, providing useful additional insight into the features of the FES community. However, labelling such data was very time consuming. Future work should look toward labelling a much larger set of users within the forum, which would allow for deeper analyses to be performed, and provide a means to evaluate belief within the groups we identified. More accurate methods for establishing belief are needed, which should draw on experts in fields such as forensic linguistics, as well as interviews with members of the community. Doing so would produce an invaluable resource for the study of belief and false information.

Further testing of language change methods

Chapters 4 and 5 highlighted various limitations of the language change methods discussed. For example, diachronic embeddings suffered from a lack of data, only made worse when splitting the data into more granular time periods or sub-groups. Changes highlighted by UFA were not very pronounced, possibly due to the shortness of the time

frame, and were not found to be useful for comparing groups in a debate setting. In VNC, the stages that were formed were not very well defined, and it is possible that tweaks to the method could ensure better clusters and more confidence in the output. ACE could be improved by increasing its explainability, for example by making it clear why language models diverge.

Future work should seek to address these limitations, and test the techniques further by applying the them to a wider range of datasets than the two used in this thesis. This will help us understand the types of community that the techniques work best on, and will show the kinds of interesting question that these methods can be used to answer. It would also be useful to test the boundaries of the methods, and establish the minimum amount of data that can yield usable results.

Multi-modal analysis

This thesis has only considered text. However, in online communities there are other important media used to communicate, such as images and video, which were beyond the scope of this thesis. Future work should aim to fill this gap. YouTube was the primary medium by which Flat Earth conspiracy theories entered the mainstream, so it is important that their language be analysed. Many of the methods described in this thesis could be applied to YouTube transcripts. This would be a logical way to extend this work. As well as video, images are another important form of communication. Analysing the contents of images and memes in conspiracy communities, and how they are shared, is an important next step in the analysis of these online movements. This has already been done to an extent for some of these communities [Zannettou et al., 2018b], but such analysis should be performed more widely.

Other Languages

This work focused on English. Future work should expand this to looking at a wide range of languages. April Fools Day is not only practised in English-speaking countries, and the corpus could be expanded to include articles written in other languages and cultures. This would go further to seeking out universal features of false information.

On a similar note, we only looked at English-speaking Flat Earth communities. In the case of Flat Earth, the bulk of interest seems to be in English speaking countries,

and the communities are so niche that there are unlikely to be significant communities in other languages, given English is often the standard language of online communication. Generally though, conspiracy theories are not a phenomenon limited to the English-speaking world. Building corpora of conspiracy communication from other languages would be a very important step in researching the universal language of conspiracy theories. This is especially relevant given that, over the last year, Covid-19 conspiracy theories have spread all around the world. Understanding how the language of these conspiracies varies from language to language and culture to culture would be an important step in fighting this harmful false information.

Bibliography

Marina Abalakina-Paap, Walter G. Stephan, Traci Craig, and W. Larry Gregory. Beliefs in conspiracies. *Political Psychology*, 20(3):637–647, 1999. doi: https://doi.org/10.1111/0162-895X.00160.

Hervé Abdi and Lynne J. Williams. Principal component analysis. *WIREs Computational Statistics*, 2(4):433–459, 2010. doi: https://doi.org/10.1002/wics.101.

Gavin Abercrombie and Riza Theresa Batista-Navarro. 'Aye' or 'No'? speech-level sentiment analysis of Hansard UK parliamentary debate transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018. URL https://www.aclweb.org/anthology/L18-1659.

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, SP '12, pages 461–475, Washington, DC, USA, 2012. IEEE Computer Society. ISBN 978-0-7695-4681-0. doi: 10.1109/SP.2012.34.

Hayri Volkan Agun and Ozgur Yilmazel. Document embedding approach for efficient authorship attribution. In *2017 2nd International Conference on Knowledge Engineering and Applications (ICKEA)*, pages 194–198, 2017. doi: 10.1109/ICKEA.2017.8169928.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1139.

Daniel Allington, Beatriz L Buarque, and Daniel Barker Flores. Antisemitic conspiracy fantasy in the age of digital media: Three 'conspiracy theorists' and their youtube audiences. *Language and Literature*, 30(1):78–102, 2021. doi: 10.1177/0963947020971997.

Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. Niche as a determinant of word fate in online groups. *PLOS ONE*, 6(5):1–12, 05 2011. doi: 10.1371/journal.pone.0019009.

Ankesh Anand, Tanmoy Chakraborty, and Noseong Park. We used neural networks to detect clickbaits: You won't believe what happened next! In Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt,

and John Tait, editors, *Advances in Information Retrieval*, pages 541–547, Cham, 2017. Springer International Publishing. ISBN 978-3-319-56608-5.

O. Aran, J. Biel, and D. Gatica-Perez. Broadcasting oneself: Visual discovery of vlogging styles. *IEEE Transactions on Multimedia*, 16(1):201–215, 2014. doi: 10.1109/TMM.2013.2284893.

Masayuki Asahara and Yuji Matsumoto. Extended models and tools for high-performance part-of-speech. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, 2000. URL https://www.aclweb.org/anthology/C00-1004.

Salvatore Attardo and Victor Raskin. Script theory revis (it) ed: Joke similarity and joke representation model. *Humor-International Journal of Humor Research*, 4(3-4): 293–348, 1991. doi: 10.1515/humr.1991.4.3-4.293.

Christoph Aymanns, Jakob Foerster, and Co-Pierre Georg. Fake news in social networks. *CoRR*, abs/1708.06233, 2017. URL http://arxiv.org/abs/1708.06233.

Mehdi Azaouzi, Delel Rhouma, and Lotfi Ben Romdhane. Community detection in large-scale social networks: state-of-the-art and future directions. *Social Network Analysis and Mining*, 9(1):23, May 2019. ISSN 1869-5469. doi: 10.1007/s13278-019-0566-x.

Jonathan Baarsch and M Emre Celebi. Investigation of internal validity measures for k-means clustering. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 14–16. sn, 2012.

Paul Baker. Times may change, but we will always have money: Diachronic variation in recent british english. *Journal of English Linguistics*, 39(1):65–88, 2011. doi: 10.1177/0075424210368368.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1389.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-main.404.

Snehasish Banerjee, Alton Y. K. Chua, and Jung-Jae Kim. Using supervised learning to classify authentic and fake online reviews. In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, IMCOM '15, pages 88:1–88:7, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3377-1. doi: 10.1145/2701126.2701130.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P14-1023.

Alexander T. J. Barron, Jenny Huang, Rebecca L. Spang, and Simon DeDeo. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1717729115.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839, May 2020. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7347.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3, 2003. URL https://www.jmlr.org/papers/volume3/tmp/bengio03a.pdf.

Shweta Bhatt, Sagar Joglekar, Shehar Bano, and Nishanth Sastry. Illuminating an ecosystem of partisan websites. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 545–554, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356404. doi: 10.1145/3184558.3188725.

Douglas Biber and Bethany Gray. The historical shift of scientific academic prose in english towards less explicit styles of expression: Writing without verbs. In *Researching Specialized Languages*, pages 11–24. John Benjamins, 2011. URL https://www.jbe-platform.com/content/books/9789027285058-scl.47.04bib.

Joan-Isaac Biel and Daniel Gatica-Perez. Vlogcast yourself: Nonverbal behavior and attention in social media. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ICMI-MLMI '10, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450304146. doi: 10.1145/1891903.1891964.

Michal Bilewicz, Mikołaj Winiewski, Mirosław Kofta, and Adrian Wójcik. Harmful ideas, the structure and consequences of anti-semitic beliefs in poland. *Political Psychology*, 34(6):821–839, 2013. doi: 10.1111/pops.12024.

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009. ISBN 0596516495.

David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 113–120, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143859.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of Machine Learning research*, 3:993–1022, 2003.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.485.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008. doi: 10.1088/1742-5468/2008/10/p10008. URL https://doi.org/10.1088/1742-5468/2008/10/p10008.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 84–89, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-4215.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D07-1090.

Lisa Branz, Patricia Brockmann, and Annika Hinze. Red is open-minded, blue is conscientious: Predicting user traits from Instagram image data. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 23–28, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.peoples-1.3.

Rachel Brazil. Fighting flat-earth theory. *Physics World*, 33(7):35–39, jul 2020. doi: 10.1088/2058-7058/33/7/34.

Vaclav Brezina. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press, 2018a. doi: 10.1017/9781316410899.

Vaclav Brezina. *Change over Time: Working Diachronic Data*, page 219–256. Cambridge University Press, 2018b. doi: 10.1017/9781316410899.008.

David A. Broniatowski, Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108(10):1378–1384, 2018. doi: 10.2105/AJPH.2018.304567. URL https://doi.org/10.2105/AJPH.2018.304567. PMID: 30138075.

Robert Brotherton, Christopher French, and Alan Pickering. Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. *Frontiers in Psychology*, 4:279, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00279.

Axel Bruns, Stephen Harrington, and Edward Hurcombe. 'corona? 5g? or both?': the dynamics of covid-19/5g conspiracy theories on facebook. *Media International Australia*, 177(1):12–29, 2020. doi: 10.1177/1329878X20946113.

Mary Bucholtz. "why be normal?": Language and identity practices in a community of nerd girls. *Language in Society*, 28(2):203–223, 1999. doi: 10.1017/S0047404599002043.

Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. Limiting the spread of misinformation in social networks. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 665–674, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963499.

Clint Burfoot and Timothy Baldwin. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 161–164, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1667583.1667633.

John Burrows. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17(3):267–287, 09 2002. ISSN 0268-1145. doi: 10.1093/llc/17.3.267.

Frances Cairncross. The cairncross review. Technical report, Department for Culture Media and Sport, HM Government, 2019. URL https://www.gov.uk/government/publications/the-cairncross-review-a-sustainable-future-for-journalism.

T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101.

Fazli Can and Jon M. Patton. Change of writing style with time. *Computers and the Humanities*, 38(1):61–82, Feb 2004. ISSN 1572-8412. doi: 10.1023/B:CHUM.0000009225.28847.77.

John R. Carlson, Joey F. George, Judee K. Burgoon, Mark Adkins, and Cindy H. White. Deception in computer-mediated communication. *Group Decision and Negotiation*, 13(1):5–28, Jan 2004. ISSN 1572-9907. doi: 10.1023/B:GRUP.0000011942.31158.d8.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 675–684, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306324. doi: 10.1145/1963405.1963500.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), December 2017. doi: 10.1145/3134666.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. Hate is not binary: Studying abusive behavior of #gamergate on twitter. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, page 65–74, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450347082. doi: 10.1145/3078714.3078721.

Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. Misleading online content: Recognizing clickbait as "false news". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, WMDD '15, pages 15–19, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3987-2. doi: 10.1145/2823465.2823467. URL `http://doi.acm.org/10.1145/2823465.2823467`.

Pride Chigwedere, George R Seage III, Sofia Gruskin, Tun-Hou Lee, and Max Essex. Estimating the lost benefits of antiretroviral drug use in south africa. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 49(4):410–415, 2008. doi: 10.1097/QAI.0b013e31818a6cd5.

Children's Commissioner. Growing up digital: A report of the growing up digital taskforce. Technical report, The Children's Commissioner, 2017. URL `https://www.childrenscommissioner.gov.uk/report/growing-up-digital/`.

Christos Christodoulopoulos, James Thorne, Andreas Vlachos, Oana Cocarascu, and Arpit Mittal, editors. *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, Online, July 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.fever-1.0`.

Jason Chuang, Margaret E. Roberts, Brandon M. Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. TopicCheck: Interactive alignment for assessing topic model stability. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 175–184, Denver, Colorado, May–June 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N15-1018.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, British Columbia, Canada, June 1989. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P89-1010.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6), 2015. doi: 10.1371/journal.pone.0128193.

Alexander Clark, Chris Fox, and Shalom Lappin. *The handbook of computational linguistics and natural language processing*. John Wiley & Sons, 2013. ISBN 9781405155816.

Michael J. Cody, Peter J. Marston, and Myrna Foster. Deception: Paralinguistic and verbal leakage. *Annals of the International Communication Association*, 8(1):464–490, 1984. doi: 10.1080/23808985.1984.11678586.

D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. doi: 10.1109/34.1000236.

Niall J. Conroy, Victoria L. Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIST '15, pages 82:1–82:4, Silver Springs, MD, USA, 2015. American Society for Information Science. ISBN 0-87715-547-X. URL http://dl.acm.org/citation.cfm?id=2857070.2857152.

Alain Couillault, Karën Fort, Gilles Adda, and Hugues de Mazancourt. Evaluating corpora documentation with regards to the ethics and big data charter. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4225–4229, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/424_Paper.pdf.

Anne Curzan. *51. Historical corpus linguistics and evidence of language change*, volume 2, pages 1091–1109. De Gruyter Mouton, 2009. doi: doi:10.1515/9783110213881.2.1091.

Andrew M. Dai, Christopher Olah, and Quoc V. Le. Document embedding with paragraph vectors, 2015.

William Dance. Disinformation online: Social media user's motivations for sharing 'fake news'. *Science in Parliament*, 75(2), November 2019. ISSN 0263-6271.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. Mark my words! linguistic style accommodation in social media. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 745–754, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306324. doi: 10.1145/1963405.1963509.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 307–318, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320351. doi: 10.1145/2488388.2488416.

S. Das Bhattacharjee, A. Talukder, and B. V. Balantrapu. Active learning based news veracity detection with feature weighting and deep-shallow fusion. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 556–565, 2017. doi: 10.1109/BigData.2017.8257971.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-3504.

D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. doi: 10.1109/TPAMI.1979.4766909.

Mark Davies. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora*, 7(2):121–157, 2012. doi: 10.3366/cor.2012.0024.

Richard Davis and Chris Proctor. Fake news, real consequences: Recruiting neural networks for the fight against fake news, 2017. URL https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2761239.pdf.

DCMS. Disinformation and fake news: Final report. Technical report, Department for Culture Media and Sport Committee, House of Commons, HMG, 2019a. URL https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf.

DCMS. Online harms white paper. Technical report, Department for Culture Media and Sport, HM Government, 2019b. URL https://www.gov.uk/government/consultations/online-harms-white-paper.

Cedric De Boom, Steven Van Canneyt, Steven Bohez, Thomas Demeester, and Bart Dhoedt. Learning semantic similarity for very short texts. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1229–1234, 2015. doi: 10.1109/ICDMW.2015.86.

Edward Dearden and Alistair Baron. Lancaster at SemEval-2018 task 3: Investigating ironic features in English tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 587–593, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-1096. URL https://www.aclweb.org/anthology/S18-1096.

Edward Dearden and Alistair Baron. Fool's errand: Looking at april fools hoaxes as disinformation through the lens of deception and humour. In *20th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2019*, April 2019a. URL https://www.research.lancs.ac.uk/portal/en/publications/fools-errand(3fb53494-6b3a-4f21-9205-d525e87fa080).html.

Edward Dearden and Alistair Baron. Fool's gold: Understanding the linguistic features of deception and humour through april fools' hoaxes. In *The 10th International Corpus Linguistics Conference, CL2019*, July 2019b. URL https://www.research.lancs.ac.uk/portal/en/publications/fools-gold(beda544a-cfa5-426d-9f82-f347b4ea4c50).html.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. doi: https://doi.org/10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9.

Stefania Degaetano-Ortlieb. Stylistic variation over 200 years of court proceedings according to gender and social class. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 1–10, New Orleans, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1601. URL https://www.aclweb.org/anthology/W18-1601.

Stefania Degaetano-Ortlieb and Elke Teich. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-4503.

Marco Del Tredici and Raquel Fernández. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1591–1603, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1135.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1210.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3): 554–559, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1517441113. URL https://www.pnas.org/content/113/3/554.

M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro, and L. de Alfaro. Automatic online fake news detection combining content and social signals. In *2018 22nd Conference of Open Innovations Association (FRUCT)*, pages 272–279, 2018. doi: 10.23919/FRUCT.2018.8468301.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1304.

Leon Derczynski and Kalina Bontcheva. Pheme: Veracity in digital social networks. In *UMAP Workshops*, 2014. URL http://www.derczynski.com/papers/synergy-workshop.pdf.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S17-2006.

Douglas C. Derrick, Thomas O. Meservy, Jeffrey L. Jenkins, Judee K. Burgoon, and Jay F. Nunamaker. Detecting deceptive chat-based communication using typing behavior and message cues. *ACM Trans. Manage. Inf. Syst.*, 4(2), August 2013. ISSN 2158-656X. doi: 10.1145/2499962.2499967.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1423.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1423.

Karen M. Douglas and Ana C. Leite. Suspicion in the workplace: Organizational conspiracy theories and work-related outcomes. *British Journal of Psychology*, 108 (3):486–506, 2017. doi: https://doi.org/10.1111/bjop.12212.

Karen M. Douglas, Robbie M. Sutton, and Aleksandra Cichocka. The psychology of conspiracy theories. *Current Directions in Psychological Science*, 26(6):538–542, 2017. doi: 10.1177/0963721417718261.

Karen M. Douglas, Joseph E. Uscinski, Robbie M. Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. Understanding conspiracy theories. *Political Psychology*, 40(S1):3–35, 2019. doi: https://doi.org/10.1111/pops.12568.

Gabriel Doyle and Michael Frank. Shared common ground influences information density in microblog texts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1587–1596, Denver, Colorado, May–June 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N15-1182.

Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74, 1993. URL https://www.aclweb.org/anthology/J93-1003.

EFAS. Action plan against disinformation. Technical report, European External Action Service, European Union, 2018. URL https://eeas.europa.eu/headquarters/headquarters-homepage/54866/action-plan-against-disinformation_en.

Steffen Eger and Alexander Mehler. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 52–58, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2009. URL https://www.aclweb.org/anthology/P16-2009.

Jacob Eisenstein. Measuring and modeling language change. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 9–14, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-5003.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. Diffusion of lexical change in social media. *PLOS ONE*, 9(11):1–13, 11 2014. doi: 10.1371/journal.pone.0113114.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 1990.

Adam Enders and Steven Smallpage. *Polls, Plots, and Party Politics: Conspiracy Theories in Contemporary America*, pages 298–318. Oxford University Press, 12 2018. ISBN 9780190844073. doi: 10.1093/oso/9780190844073.003.0020.

Stefan Evert. *The statistics of word cooccurrences: word pairs and collocations*. PhD thesis, University of Stuttgart, 2005. URL http://dx.doi.org/10.18419/opus-2556.

Stefan Evert, Thomas Proisl, Fotis Jannidis, Isabella Reger, Steffen Pielström, Christof Schöch, and Thorsten Vitt. Understanding and explaining delta measures for authorship attribution. *Digital Scholarship in the Humanities*, 2017.

Kate Faasse, Casey J. Chatman, and Leslie R. Martin. A comparison of language use in pro- and anti-vaccination comments in response to a high profile facebook post,. *Vaccine*, 34(47):5808–5814, 2016. ISSN 0264-410X. doi: https://doi.org/10.1016/j.vaccine.2016.09.029.

Don Fallis. What Is Disinformation? *Library Trends*, 63(3):401–426, 2015. ISSN 1559-0682. doi: 10.1353/lib.2015.0014.

Hao Fang, Hao Cheng, and Mari Ostendorf. Learning latent local conversation modes for predicting comment endorsement in online discussions. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 55–64, Austin, TX, USA, November 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W16-6209.

Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 171–175, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390665.2390708.

William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California, June 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N16-1138.

David Fifield, Torbjørn Follan, and Emil Lunde. Unsupervised authorship attribution. *arXiv preprint arXiv:1503.07613*, 2015.

John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. The royal society corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 794–802, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.99.

Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3): 221–233, 1948. doi: 10.1037/h0057532. URL https://doi.org/10.1037/h0057532.

Richard S Forsyth. Stylochronometry with substrings, or: A poet young and old. *Literary and Linguistic Computing*, 1999.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2009.11.002.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1): 32–40, 1975. doi: 10.1109/TIT.1975.1055330.

Dana Gablasova, Vaclav Brezina, and Tony McEnery. Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning*, 67(S1):130–154, 2017. doi: https://doi.org/10.1111/lang.12226.

Costas Gabrielatos, Tony McEnery, Peter J. Diggle, and Paul Baker. The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics*, 17(2):151–175, 2012. ISSN 1384-6655. doi: https://doi.org/10.1075/ijcl.17.2.01gab.

Devin Gaffney and J. Nathan Matias. Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus. *PLOS ONE*, 13(7): 1–13, 07 2018. doi: 10.1371/journal.pone.0200162.

Natasha Galliford and Adrian Furnham. Individual difference factors and beliefs in medical and political conspiracy theories. *Scandinavian Journal of Psychology*, 58 (5):422–428, 2017. doi: https://doi.org/10.1111/sjop.12382.

David Garcia, Ingmar Weber, and Venkata Garimella. Gender asymmetries in reality and fiction: The bechdel test of social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), May 2014. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/14522`.

Matt Garley and Julia Hockenmaier. Beefmoves: Dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 135–139, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P12-2027`.

Roger Garside. The claws word-tagging system. *The Computational analysis of English: A corpus-based approach. London: Longman*, pages 30–41, 1987.

Aaron Gerow, Yuening Hu, Jordan Boyd-Graber, David M. Blei, and James A. Evans. Measuring discursive influence across scholarship. *Proceedings of the National Academy of Sciences*, 115(13):3308–3313, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1719792115.

Sean M. Gerrish and David M. Blei. A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 375–382, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,

pages 4952–4957, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1537.

Mathew Gillings. A topic modelling analysis of the Lancaster corpus on climate change (British) and a methodological evaluation of topic model interpretation. Master's thesis, Lancaster University, 2016. URL https://www.academia.edu/33730081/A_topic_modelling_analysis_of_the_Lancaster_Corpus_on_Climate_Change_British_and_a_methodological_evaluation_of_topic_model_interpretation.

Mathew Gillings. *A corpus-based investigation into verbal cues to deception and their sociolinguistic distribution*. PhD thesis, Lancaster University, 2021. Embargoed for 5 years.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.365.

M. Glenski, C. Pennycuff, and T. Weninger. Consumers and curators: Browsing and voting patterns on reddit. *IEEE Transactions on Computational Social Systems*, 4(4): 196–206, 2017. doi: 10.1109/TCSS.2017.2742242.

Maria Glenski and Tim Weninger. Predicting user-interactions on reddit. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM '17, page 609–612, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349932. doi: 10.1145/3110025.3120993.

Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. The social dynamics of language change in online networks. In Emma Spiro and Yong-Yeol Ahn, editors, *Social Informatics*, pages 41–57, Cham, 2016. Springer International Publishing. ISBN 978-3-319-47880-7.

Ted Goertzel. Belief in conspiracy theories. *Political Psychology*, 15(4):731–742, 2021/03/29/ 1994. doi: 10.2307/3791630. Full publication date: Dec., 1994.

Ted Goertzel. Conspiracy theories in science. *EMBO reports*, 11(7):493–499, 2010. doi: https://doi.org/10.1038/embor.2010.84.

Yoav Goldberg. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309, 2017. doi: 10.2200/S00762ED1V01Y201703HLT037.

Scott A. Golder and Michael W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011. ISSN 0036-8075. doi: 10.1126/science.1202775.

Nataša Golo and Serge Galam. Conspiratorial beliefs observed through entropy principles. *Entropy*, 17(12):5611–5634, Aug 2015. ISSN 1099-4300. doi: 10.3390/e17085611.

Wael H Gomaa and Aly A Fahmy. A survey of text similarity approaches. *International Journal of Computer Applications*, 2013.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.51.

Stefan Th. Gries. *Statistics for Linguistics with R*. De Gruyter Mouton, 2013. doi: doi:10.1515/9783110307474.

Stefan Th. Gries and Martin Hilpert. The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora*, 3(1):59–81, 2008. doi: 10.3366/E1749503208000075.

Jack Grieve. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270, 07 2007. ISSN 0268-1145. doi: 10.1093/llc/fqm020. URL https://doi.org/10.1093/llc/fqm020.

Jack Grieve, Andrea Nini, and Diansheng Guo. Analyzing lexical emergence in modern american english online. *English Language and Linguistics*, 21(1):99–127, 2017. doi: 10.1017/S1360674316000113.

Monika Grzesiak-Feldman. The effect of high-anxiety situations on conspiracy thinking. *Current Psychology*, 32(1):100–118, Mar 2013. ISSN 1936-4733. doi: 10.1007/s12144-013-9165-6.

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of false information detection on social media: New perspectives and trends. *ACM Comput. Surv.*, 53(4), July 2020. ISSN 0360-0300. doi: 10.1145/3393880.

Fangjian Guo, Charles Blundell, Hanna Wallach, and Katherine Heller. The Bayesian Echo Chamber: Modeling Social Influence via Linguistic Accommodation. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 315–323, San Diego, California, USA, 09–12 May 2015. PMLR. URL http://proceedings.mlr.press/v38/guo15.html.

Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 729–736, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320382. doi: 10.1145/2487788.2488033.

Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. *TweetCred: Real-Time Credibility Assessment of Content on Twitter*, pages 228–243. Springer International Publishing, Cham, 2014. ISBN 978-3-319-13734-6. doi: 10.1007/978-3-319-13734-6_16.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1): 29–48, 2008. doi: https://doi.org/10.1348/000711006X126600.

David Hall, Daniel Jurafsky, and Christopher D. Manning. Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, page 363–371, USA, 2008. Association for Computational Linguistics.

Oren Halvani, Christian Winter, and Anika Pflug. Authorship verification for different languages, genres and topics. *Digital Investigation*, 16:S33–S43, 2016. ISSN 1742-2876. doi: 10.1016/j.diin.2016.01.006. DFRWS 2016 Europe.

William Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. Loyalty in online communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), May 2017. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14972.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas, November 2016a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D16-1229.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August 2016b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P16-1141.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas, November 2016c. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D16-1229.

Jeffrey Hancock, Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45:1–23, 01 2007. doi: 10.1080/01638530701739181.

Andrew Hardie. Cqpweb combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409, 2012. ISSN 1384-6655. doi: doi:10.1075/ijcl.17.3.04har.

Alice C. Harris. Revisiting anaphoric islands. *Language*, 82(1):114–130, 2006. ISSN 00978507, 15350665. URL http://www.jstor.org/stable/4490086.

Zellig S Harris. Distributional structure. *Word*, 10, 1954.

James Hartley. Is time up for the flesch measure of reading ease? *Scientometrics*, 107 (3):1523–1526, Jun 2016. ISSN 1588-2861. doi: 10.1007/s11192-016-1920-7. URL https://doi.org/10.1007/s11192-016-1920-7.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450337946. doi: 10.1145/2806416.2806652.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. Claimbuster: The first-ever end-to-end fact-checking system. *Proc. VLDB Endow.*, 10(12): 1945–1948, August 2017. ISSN 2150-8097. doi: 10.14778/3137765.3137815.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, 2010. doi: 10.1017/S0140525X0999152X.

Peter Hernon. Disinformation and misinformation through the internet: Findings of an exploratory study. *Government Information Quarterly*, 12(2):133–139, 1995. ISSN 0740-624X. doi: https://doi.org/10.1016/0740-624X(95)90052-7.

Jack Hessel, Chenhao Tan, and Lillian Lee. Science, askscience, and badscience: On the coexistence of highly related communities. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1), Mar. 2016. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/14739`.

Martin Hilpert and Stefan Th. Gries. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4):385–401, 09 2008. ISSN 0268-1145. doi: 10.1093/llc/fqn012.

Martin Hilpert and Stefan Th. Gries. *Quantitative approaches to diachronic corpus linguistics*, page 36–53. Cambridge Handbooks in Language and Linguistics. Cambridge University Press, 2016. doi: 10.1017/CBO9781139600231.003.

Gabriel Hine, Jeremiah Onaolapo, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Riginos Samaras, Gianluca Stringhini, and Jeremy Blackburn. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), May 2017. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/14893`.

Klaus Hofmann, Anna Marakasova, Andreas Baumann, Julia Neidhardt, and Tanja Wissik. Comparing lexical usage in political discourse across diachronic corpora. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 58–65, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-47-4. URL `https://www.aclweb.org/anthology/2020.parlaclarin-1.11`.

Robert Hogg, Busisiwe Nkala, Janan Dietrich, Alexandra Collins, Kalysha Closson, Zishan Cui, Steve Kanters, Jason Chia, Bernard Barhafuma, Alexis Palmer, Angela Kaida, Glenda Gray, and Cari Miller. Conspiracy beliefs and knowledge about

hiv origins among adolescents in soweto, south africa. *PLOS ONE*, 12(2):1–9, 02 2017. doi: 10.1371/journal.pone.0165087. URL https://doi.org/10.1371/journal.pone.0165087.

Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017a.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017b.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

David L. Hoover. Corpus stylistics, stylometry, and the styles of henry james. *Style*, 41(2):174–203, 2007. ISSN 00394238, 23746629. URL http://www.jstor.org/stable/10.5325/style.41.2.174.

B. D. Horne, S. Adali, and S. Sikdar. Identifying the social signals that drive online discussions: A case study of reddit communities. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9, 2017. doi: 10.1109/ICCCN.2017.8038388.

Benjamin Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), May 2017. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14976.

Benjamin Horne and Mauricio Gruppi. NELA-GT-2020. doi.org/10.7910/DVN/CHMUYZ, 2021.

Dirk Hovy. The enemy in your own camp: How well can we detect statistically-generated fake reviews – an adversarial study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 351–356, Berlin, Germany, August 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P16-2057.

Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China, July 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P15-2079.

Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P16-2096.

Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Social spammer detection with sentiment information. In *Proceedings of the 2014 IEEE International Conference on Data Mining*, ICDM '14, page 180–189, USA, 2014. IEEE Computer Society. ISBN 9781479943029. doi: 10.1109/ICDM.2014.141.

Faliang Huang, Chaoxiong Li, and Li Lin. Identifying gender of microblog users based on message mining. In Feifei Li, Guoliang Li, Seung-won Hwang, Bin Yao, and Zhenjie Zhang, editors, *Web-Age Information Management*, pages 488–493, Cham, 2014. Springer International Publishing. ISBN 978-3-319-08010-9.

Roland Imhoff and Pia Lamberty. How paranoid are conspiracy believers? toward a more fine-grained understanding of the connect and disconnect between paranoia and belief in conspiracy theories. *European Journal of Social Psychology*, 48(7):909–926, 2018. doi: https://doi.org/10.1002/ejsp.2494.

Jun Ito, Jing Song, Hiroyuki Toda, Yoshimasa Koike, and Satoshi Oyama. Assessment of tweet credibility with lda features. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 953–958, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334730. doi: 10.1145/2740908.2742569.

Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes. Affect, Not Ideology: A Social Identity Perspective on Polarization. *Public Opinion Quarterly*, 76(3):405–431, 09 2012. ISSN 0033-362X. doi: 10.1093/poq/nfs038.

Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. Talking to the crowd: What do people react to in online discussions? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2026–2031, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D15-1239`.

Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini. Measuring gender bias in news images. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 893–898, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334730. doi: 10.1145/2740908.2742007.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. Novel visual and statistical image features for microblogs news verification. *Trans. Multi.*, 19 (3):598–608, March 2017. ISSN 1520-9210. doi: 10.1109/TMM.2016.2617078.

Matthew L. Jockers and Daniela M. Witten. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2):215–223, 04 2010. ISSN 0268-1145. doi: 10.1093/llc/fqq001.

Neil F. Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. The online competition between pro- and anti-vaccination views. *Nature*, 582(7811):230–233, Jun 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2281-1.

Daniel Jolley and Karen M. Douglas. The social consequences of conspiracism: Exposure to conspiracy theories decreases intentions to engage in politics and to reduce one's carbon footprint. *British Journal of Psychology*, 105(1):35–56, 2014a. doi: https://doi.org/10.1111/bjop.12018.

Daniel Jolley and Karen M. Douglas. The effects of anti-vaccine conspiracy theories on vaccination intentions. *PLOS ONE*, 9(2):1–9, 02 2014b. doi: 10.1371/journal.pone.0089177.

Kayla N. Jordan, Joanna Sterling, James W. Pennebaker, and Ryan L. Boyd. Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences*, 116, 2019. doi: https://doi.org/10.1073/pnas.1811987116.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China, July 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W15-4302.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., 2 edition, 2009. ISBN 0131873210.

Sanjay Ram Kairam, Dan J. Wang, and Jure Leskovec. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, page 673–682, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450307475. doi: 10.1145/2124295.2124374.

Hamid Karimi, Proteek Roy, Sari Saba-Sadiya, and Jiliang Tang. Multi-source multi-class fake news detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1546–1557, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1131.

Anna Kata. A postmodern pandora's box: Anti-vaccination misinformation on the internet. *Vaccine*, 28(7):1709–1716, 2010. ISSN 0264-410X. doi: https://doi.org/10.1016/j.vaccine.2009.12.022.

Anna Kata. Anti-vaccine activists, web 2.0, and the postmodern paradigm – an overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine*, 30(25):3778–3789, 2012. ISSN 0264-410X. doi: https://doi.org/10.1016/j.vaccine.2011.11.112. Special Issue: The Role of Internet Use in Vaccination Decisions.

P. S. Keila and D. B. Skillicorn. Detecting unusual email communication. In *Proceedings of the 2005 Conference of the Centre for Advanced Studies on Collaborative Research*, CASCON '05, page 117–125. IBM Press, 2005. URL https://dl.acm.org/doi/abs/10.5555/1105634.1105643.

Daniel Kershaw, Matthew Rowe, Anastasios Noulas, and Patrick Stacey. Birds of a feather talk together: user influence on language adoption. In *Proceedings of the*

50th Hawaii International Conference on System Sciences, pages 1851–1860. IEEE, 2017. doi: 10.24251/HICSS.2017.225.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, page 2915–2921, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313552.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S19-2145.

Adam Kilgarriff. Comparing corpora. *International Journal of Corpus Linguistics*, 6 (1):97–133, 2001. ISSN 1384-6655. doi: https://doi.org/10.1075/ijcl.6.1.05kil.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. The sketch engine. *Information Technology*, 105, 2004.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA, jun 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W14-2517.

Carmen Klaussner and Carl Vogel. Stylochronometry: Timeline prediction in stylometric analysis. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 2015.

Carmen Klaussner and Carl Vogel. A diachronic corpus for literary style analysis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018a. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1552.

Carmen Klaussner and Carl Vogel. Temporal predictive regression models for linguistic style analysis. *Journal of Language Modelling*, 2018b. doi: 10.15398/jlm.v6i1.177.

Bennett Kleinberg, Maximilian Mozes, and Isabelle van der Vegt. Identifying the sentiment styles of YouTube's vloggers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3581–3590, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1394.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S17-2083.

Miroslaw Kofta, , and Grzegorz Sedek. Conspiracy stereotypes of jews during systemic transformation in poland. *International Journal of Sociology*, 35(1):40–64, 2005. doi: 10.1080/00207659.2005.11043142.

Mirosław Kofta, Wiktor Soral, and Michał Bilewicz. What breeds conspiracy antisemitism? the role of political uncontrollability and uncertainty in the belief in jewish conspiracy. *Journal of Personality and Social Psychology*, 2020. doi: 10.1037/pspa0000183.

Farshad Kooti, Haeryun Yang, Meeyoung Cha, Krishna Gummadi, and Winter Mason. The emergence of conventions in online social networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1), May 2012. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14267.

Dimitrios Kouzis-Loukas. *Learning Scrapy*. Packt Publishing Ltd, 2016.

Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data*, 3(1), March 2009. ISSN 1556-4681. doi: 10.1145/1497577.1497578.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in Instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-1469.

Rohan Kshirsagar, Robert Morris, and Samuel Bowman. Detecting and explaining crisis. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 66–73, Vancouver, BC, August 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-3108.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741627. URL https://doi.org/10.1145/2736277.2741627.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22, 1951. URL https://www.jstor.org/stable/2236703.

Srijan Kumar and Neil Shah. False information on web and social media: A survey, 2018.

Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 933–943, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. doi: 10.1145/3178876.3186141.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France, 07–09 Jul 2015. PMLR. URL `http://proceedings.mlr.press/v37/kusnerb15.html`.

Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. Tracing armed conflicts with diachronic word embedding models. In *Proceedings of the Events and Stories in the News Workshop*, pages 31–36, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W17-2705`.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/C18-1117`.

S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108. IEEE, 2013. doi: 10.1109/ICDM.2013.61.

Merja Kytö. Corpora and historical linguistics. *Revista Brasileira de Linguística Aplicada*, 11(2):417–457, 2011.

William Labov. *The social stratification of English in New York city.* ERIC, 1966.

William Labov. *Principles of linguistic change, volume 3: Cognitive and cultural factors*. John Wiley & Sons, 2011.

Evan E Laine and Raju Parakkal. National security, personal insecurity, and political conspiracies: The persistence of americans' beliefs in 9/11 conspiracy theories. *IUP Journal of International Relations*, 11(3), 2017.

R. Lambiotte and M. Kosinski. Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12):1934–1939, 2014. doi: 10.1109/JPROC.2014.2359054.

Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: A comparative analysis. *Phys. Rev. E*, 80:056117, Nov 2009. doi: 10.1103/PhysRevE.80.056117.

Asheley Landrum and Alex Olshansky. Third-person perceptions and calls for censorship of flat earth videos on youtube. *Media and Communication*, 8(2):387–400, 2020. ISSN 2183-2439. doi: 10.17645/mac.v8i2.2853.

Asheley R. Landrum and Alex Olshansky. 2017 flat earth conference interviews, 2019. accessed: 2021/12/18.

Asheley R. Landrum, Alex Olshansky, and Othello Richards. Differential susceptibility to misleading flat earth arguments on youtube. *Media Psychology*, 24(1):136–165, 2021. doi: 10.1080/15213269.2019.1669461.

Benjamin E Lauderdale and Alexander Herzog. Measuring political positions from legislative speech. *Political Analysis*, 24, 2016. URL https://www.jstor.org/stable/26349743.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/le14.html.

Jong Gun Lee, Sue Moon, and Kavé Salamatian. Modeling and predicting the popularity of online contents with cox proportional hazard regression model. *Neurocomputing*, 76(1):134–145, 2012. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2011.04.040. Seventh International Symposium on Neural Networks (ISNN 2010) Advances in Web Intelligence.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.fever-1.5.

Geoffrey Leech, Marianne Hundt, Christian Mair, and Nicholas Smith. *Change in Contemporary English: A Grammatical Study*. Studies in English Language. Cambridge University Press, 2009. doi: 10.1017/CBO9780511642210.

Geoffrey Leech, Nicholas Smith, and Paul Rayson. *English style on the move: variation and change in stylistic norms in the twentieth century*, pages 69–98. Brill, Leiden, The Netherlands, 2012. ISBN 9789401207935. doi: https://doi.org/10.1163/9789401207935_006.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1011. URL https://www.aclweb.org/anthology/D16-1011.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 497–506, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584959. doi: 10.1145/1557019.1557077.

Stephan Lewandowsky, Klaus Oberauer, and Gilles E. Gignac. Nasa faked the moon landing—therefore, (climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological Science*, 24(5):622–633, 2013. doi: 10.1177/0956797612457686. PMID: 23531484.

Heidi Oi-Yee Li, Adrian Bailey, David Huynh, and James Chan. YouTube as a source of information on COVID-19: a pandemic of misinformation? *BMJ Global Health*, 5(5), 2020a. URL https://gh.bmj.com/content/5/5/e002604.

Quanzhi Li, Qiong Zhang, and Luo Si. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, Florence, Italy, July 2019a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1113.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. Is word segmentation necessary for deep learning of Chinese representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3242–3252, Florence, Italy, July 2019b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1314.

Yunyao Li, Tyrone Grandison, Patricia Silveyra, Ali Douraghy, Xinyu Guan, Thomas Kieselbach, Chengkai Li, and Haiqi Zhang. Jennifer for COVID-19: An NLP-powered chatbot built for the people and by the people to combat misinformation. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.nlpcovid19-acl.9.

Wee Yong Lim, Mong Li Lee, and Wynne Hsu. Ifact: An interactive framework to assess claims from tweets. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 787–796, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450349185. doi: 10.1145/3132847.3132995.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1867–1870, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450337946. doi: 10.1145/2806416.2806651.

Y. Liu and S. Xu. Detecting rumors through modeling information propagation networks in a social media environment. *IEEE Transactions on Computational Social Systems*, 3(2):46–62, 2016. doi: 10.1109/TCSS.2016.2612980.

Yang Liu and Yi-Fang Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11268.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-2043.

Cristian Lumezanu, Nick Feamster, and Hans Klein. #bias: Measuring the tweeting behavior of propagandists. In *International AAAI Conference on Web and Social Media*. AAAI, 2012. URL https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4588/4985.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1751–1754, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450337946. doi: 10.1145/2806416.2806607.

Christian Mair. *52. Corpora and the study of recent change in language*, volume 2, pages 1109–1125. De Gruyter Mouton, 2009. doi: doi:10.1515/9783110213881.2.1109.

H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1): 50 – 60, 1947. doi: 10.1214/aoms/1177730491.

Christopher D. Manning. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 171–189, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-19400-9.

Alexios Mantzarlis. Will verification kill fact-checking?, 2015. URL https://www.poynter.org/fact-checking/2015/will-verification-kill-fact-checking/. accessed: 2021/12/18.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1062.

Andrey Markov. *Theory of Algorithms*. Academy of Sciences of the USSR, 1954.

Ilia Markov, Helena Gómez-Adorno, Juan-Pablo Posadas-Durán, Grigori Sidorov, and Alexander Gelbukh. Author profiling with doc2vec neural network-based document embeddings. In Obdulia Pichardo-Lagunas and Sabino Miranda-Jiménez, editors, *Advances in Soft Computing*, pages 117–131, Cham, 2017. Springer International Publishing. ISBN 978-3-319-62428-0.

David M. Markowitz and Jeffrey T. Hancock. Linguistic traces of a scientific fraud: The case of diederik stapel. *PLOS ONE*, 9(8):1–5, 08 2014. doi: 10.1371/journal.pone.0105937. URL https://doi.org/10.1371/journal.pone.0105937.

D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 International Seminar on Application for Technology of Information and Communication*, pages 533–538, 2018. doi: 10.1109/ISEMANTIC.2018.8549751.

Jaume Masip, Maria Bethencourt, Guadalupe Lucas, MIRIAM SÁNCHEZ-SAN SEGUNDO, and Carmen Herrero. Deception detection from written accounts. *Scandinavian Journal of Psychology*, 53(2):103–111, 2012. doi: 10.1111/j.1467-9450.2011.00931.x.

Anthony McEnery and Helen Baker. *Corpus Linguistics and 17th-century prostitution: computational linguistics and history*. Bloomsbury Academic, 2016. ISBN 9781472506092.

Tony McEnery and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2011. doi: 10.1017/CBO9780511981395.

Tony McEnery, Vaclav Brezina, and Helen Baker. Usage fluctuation analysis: A new way of analysing shifts in historical discourse. *International Journal of Corpus Linguistics*, 24(4):413–444, 2019. ISSN 1384-6655. doi: https://doi.org/10.1075/ijcl.18096.mce. URL https://www.jbe-platform.com/content/journals/10.1075/ijcl.18096.mce.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.nlpcovid19-acl.17.

Alexey N. Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. The anatomy of reddit: An overview of academic research. In Fakhteh Ghanbarnejad, Rishiraj Saha Roy, Fariba Karimi, Jean-Charles Delvenne, and Bivas Mitra, editors, *Dynamics On and Of Complex Networks III*, pages 183–204, Cham, 2019. Springer International Publishing. ISBN 978-3-030-14683-2.

Panagiotis Takis Metaxas. Web spam, social propaganda and the evolution of search engine rankings. In Joaquim Cordeiro, Joséand Filipe, editor, *Web Information Systems and Technologies*, pages 170–182, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-12436-5. URL https://link.springer.com/chapter/10.1007/978-3-642-12436-5_13.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011. ISSN 0036-8075. doi: 10.1126/science.1199644.

Margot Mieskes. A quantitative study of data in the NLP community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-1603.

Rada Mihalcea and Stephen Pulman. Characterizing humour: An exploration of features in humorous texts. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 337–347, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-70939-8. URL https://link.springer.com/chapter/10.1007/978-3-540-70939-8_30.

Rada Mihalcea and Carlo Strapparava. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/H05-1067.

Márton Miháltz, Tamás Váradi, István Csertő, Éva Fülöp, Tibor Pólya, and Pál Kővágó. Beyond sentiment: Social psychological analysis of political Facebook comments in Hungary. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 127–133, Lisboa, Portugal, September 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W15-2918.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*, 2013. URL http://arxiv.org/abs/1301.3781.

Tanushree Mitra and Eric Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1), Apr. 2015. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14625.

Shaheed Mohammed. Conspiracy theories and flat-earth videos on youtube. *The Journal of Social Media in Society*, 8(2):84–102, 2019. ISSN 2325-503x. URL https://www.thejsms.org/index.php/TSMRI/article/view/527.

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan, Mitesh M. Khapra, Balaji Vasan Srinivasan, and Balaraman Ravindran. Towards transparent and explainable attention models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4206–4216, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.387.

Luis Gerardo Mojica de la Vega and Vincent Ng. Modeling trolling in social media conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1585.

Sandra Mollin. The Hansard hazard: gauging the accuracy of British parliamentary transcripts. *Corpora*, 2, 2007. doi: https://doi.org/10.3366/cor.2007.2.2.187.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403, 2017. doi: 10.1093/pan/mpn018.

Roser Morante, Chantal van Son, Isa Maks, and Piek Vossen. Annotating perspectives on vaccination. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4964–4973, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.611.

Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1), Jun. 2013. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14401.

Rachel R. Mourão and Craig T. Robertson. Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information. *Journalism Studies*, 20(14):2077–2095, 2019. doi: 10.1080/1461670X.2019.1566871.

Daniel Moyer, Samuel Carson, Thayne Dye, Richard Carson, and David Goldbaum. Determining the influence of reddit posts on wikipedia pageviews. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1), Apr. 2015. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14700.

Damian Mrowca, Elias Wang, and Atli Kosson. Stance detection for fake news identification, 2017. URL https://eliaszwang.com/project/stance-detection/stance-detection.pdf.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D15-1272.

Eni Mustafaraj and Panagiotis Takis Metaxas. The fake news spreading plague: Was it preventable? In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 235–239, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4896-6. doi: 10.1145/3091478.3091523.

Vineeth G. Nair. *Getting Started with Beautiful Soup*. Packt, 2014. ISBN 9781783289554. URL https://www.packtpub.com/product/getting-started-with-beautiful-soup/9781783289554.

Ndapandula Nakashole and Tom M. Mitchell. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019, Baltimore,

Maryland, June 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P14-1095.

Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy*, pages 300–314, 2012. doi: 10.1109/SP.2012.46.

Gonzalo Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33:31–88, 2001. ISSN 0360-0300. doi: 10.1145/375360.375365.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, 50(6), November 2017. ISSN 0360-0300. doi: 10.1145/3132039.

Edward Newell, David Jurgens, Haji Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. User migration in online social networks: A case study on reddit during a period of community unrest. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1), Mar. 2016. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14750.

Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675, May 2003. ISSN 0146-1672. doi: 10.1177/0146167203029005010.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 849–856, Cambridge, MA, USA, 2001. MIT Press.

Dong Nguyen and Carolyn P. Rosé. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 76–85, Portland, Oregon, June 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-0710.

Nam P. Nguyen, Guanhua Yan, My T. Thai, and Stephan Eidenbenz. Containment of misinformation spread in online social networks. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 213–222, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1228-8. doi: 10.1145/2380718.2380746.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah A Cai, Jennifer E Midberry, and Yuanxin Wang. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95(3):381–421, 2014. doi: 10.1007/s10994-013-5417-9.

Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175–220, 1998. ISSN 1089-2680. doi: 10.1037/1089-2680.2.2.175.

Kate G. Niederhoffer and James W. Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360, Dec 2002. ISSN 0261-927X. doi: 10.1177/026192702237953. URL https://doi.org/10.1177/026192702237953.

Frank Nielsen. *Hierarchical Clustering*, pages 195–211. Springer International Publishing, Cham, 2016. ISBN 978-3-319-21903-5. doi: 10.1007/978-3-319-21903-5_8. URL https://doi.org/10.1007/978-3-319-21903-5_8.

ODNI. Assessing Russian Activities and Intentions in Recent US Elections. Technical report, Office of the Director of National Intelligence, 01 2017. URL https://www.dni.gov/files/documents/ICA_2017_01.pdf.

J. Eric Oliver and Thomas J. Wood. Conspiracy theories and the paranoid style(s) of mass opinion. *American Journal of Political Science*, 58(4):952–966, 2014. doi: https://doi.org/10.1111/ajps.12084.

Alex Olshansky, Robert M. Peaslee, and Asheley R. Landrum. Flat-smacked! converting to flat eartherism. *Journal of Media and Religion*, 19(2):46–59, 2020. doi: 10.1080/15348423.2020.1774257.

Alex Olshansky et al. Conspiracy theorizing and religious motivated reasoning: Why the earth 'must'be flat. Master's thesis, Texas Tech University, 2018.

Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.747.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 309–319, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9. URL http://dl.acm.org/citation.cfm?id=2002472.2002512.

Jan Overgoor, Bogdan State, and Lada A. Adamic. The structure of u.s. college networks on facebook. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):499–510, May 2020. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7318.

John C. Paolillo. The flat earth phenomenon on youtube. *First Monday*, 23(12), Dec. 2018. doi: 10.5210/fm.v23i12.8251.

Antonis Papasavva, Jeremy Blackburn, Gianluca Stringhini, Savvas Zannettou, and Emiliano De Cristofaro. "is it a qoincidence?": An exploratory study of qanon on voat. In *Proceedings of the Web Conference 2021*, WWW '21, page 460–471, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3450036. URL https://doi.org/10.1145/3442381.3450036.

Fuchun Peng, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. Language independent authorship attribution using character level language models. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, page 267–274, USA, 2003. Association for Computational Linguistics. ISBN 1333567890. doi: 10.3115/1067807.1067843.

James W Pennebaker. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Publishing USA, 2013. ISBN 1608194965.

James W. Pennebaker and Lori D. Stone. Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2):291–301, 2003. doi: 10.1037/0022-3514.85.2.291. URL https://doi.org/10.1037/0022-3514.85.2.291.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2015, 2001. URL http://downloads.liwc.net.s3.amazonaws.com/LIWC2015_OperatorManual.pdf. accessed: 2021/12/18.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D14-1162.

Florent Perek. Vector spaces for historical linguistics: Using distributional semantics to study syntactic productivity in diachrony. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 309–314, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P14-2051.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1287.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-1202.

Andrew Peterson and Arthur Spirling. Classification accuracy as a substantive quantity of interest: Measuring polarization in Westminster systems. *Political Analysis*, 26:120–128, 2018. doi: https://doi.org/10.1017/pan.2017.39.

Sangita R Pillay and Thamar Solorio. Authorship attribution of web forum posts. In *2010 eCrime Researchers Summit*, pages 1–7. IEEE, 2010. doi: 10.1109/ecrime.2010.5706693.

Dean Pomerleau and Delip Rao. Fake news challenge, 2017. URL http://www.fakenewschallenge.org. accessed: 2021/12/18.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 155–158, Republic and Canton of Geneva, CHE, 2018a. International World Wide Web Conferences Steering Committee. ISBN 9781450356404. doi: 10.1145/3184558.3186967.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1003.

Octavian Popescu and Carlo Strapparava. Behind the times: Detecting epoch changes using large corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 347–355, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I13-1040.

Octavian Popescu and Carlo Strapparava. Time corpora: Epochs, opinions and changes. *Knowledge-Based Systems*, 69:3–13, 2014. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2014.04.029.

Martin F Porter. An algorithm for suffix stripping. *Program*, 1980. URL https://doi.org/10.1108/eb046814.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P18-1022.

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1589–1599, USA, 2011. Association for Computational Linguistics. ISBN 9781937284114.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July 2020a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-demos.14.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July 2020b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-demos.14.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1317.

Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: Mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 249–252, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0637-9. doi: 10.1145/1963192.1963301.

Paul Rayson. From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549, 2008. ISSN 1384-6655. doi: https://doi.org/10.1075/ijcl.13.4.06ray.

Paul Rayson and Roger Garside. Comparing corpora using frequency profiling. In *The Workshop on Comparing Corpora*, pages 1–6, Hong Kong, China, October 2000. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W00-0901.

Paul Rayson, Andrew Wilson, and Geoffrey Leech. Grammatical word class variation within the british national corpus sampler. *Language and Computers*, 36, 2002.

Paul Rayson, Dawn Archer, Scott Piao, and Anthony M McEnery. The ucrel semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop*, pages 7–12, Lisbon, Portugal, 2004. URL https://eprints.lancs.ac.uk/id/eprint/1783/.

Raquel Recuero, Felipe Bonow Soares, and Anatoliy Gruzd. Hyperpartisanship, disinformation and political conversations on twitter: The brazilian presidential election of 2018. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):569–578, May 2020. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7324.

T Raghunadha Reddy, B Vishnu Vardhan, and P Vijaypal Reddy. A survey on authorship profiling techniques. *International Journal of Applied Engineering Research*, 11(5): 3092–3102, 2016.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. European Language Resources Association, 2010.

Julio C. S. Reis, Philipe Melo, Kiran Garimella, Jussara M. Almeida, Dean Eckles, and Fabrício Benevenuto. A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):903–908, May 2020. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7356.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. Humor in the blogosphere: First clues for a verbal humor taxonomy. *Journal of Intelligent Systems*, 18(4):311–332, 2009. doi: doi:10.1515/JISYS.2009.18.4.311. URL https://doi.org/10.1515/JISYS.2009.18.4.311.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12, 2012.

Delel Rhouma and Lotfi Ben Romdhane. An efficient algorithm for community mining with overlap in social networks. *Expert Systems with Applications*, 41(9):4309–4321, 2014. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2014.01.002.

CM Rivers and BL Lewis. Ethical research standards in a world of big data. *F1000Research*, 3(38), 2014. doi: 10.12688/f1000research.3-38.v2.

Peter Ronhovde and Zohar Nussinov. Multiresolution community detection for megascale networks by information-based replica correlations. *Phys. Rev. E*, 80: 016109, Jul 2009. doi: 10.1103/PhysRevE.80.016109. URL https://link.aps.org/doi/10.1103/PhysRevE.80.016109.

Jon Roozenbeek and Sander van der Linden. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 5(1), 2019. doi: 10.1057/s41599-019-0279-9.

Jonathan Rose. Brexit, trump, and post-truth politics. *Public Integrity*, 19(6):555–558, 2017. doi: 10.1080/10999922.2017.1285540.

Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105 (4):1118–1123, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0706851105.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7.

Samuel B. Rowbotham. *Zetetic astronomy. Earth not a globe! an experimental inquiry into the true figure of the earth, by 'Parallax'.* 1865. URL https://books.google.co.uk/books?id=oTUDAAAAQAAJ.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California, June 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W16-0802.

Victoria L Rubin, Niall J Conroy, and Yimin Chen. Towards news verification: Deception detection methods for news discourse. In *Hawaii International Conference on System Sciences*, Hawaii, USA, 2015. IEEE Computer Society. doi: 10.13140/2.1.4822.8166.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 2000.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-5004.

Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA, 1986. ISBN 0070544840. URL https://dl.acm.org/doi/10.5555/576628.

Mattia Samory and Tanushree Mitra. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, Jun 2018.

Giovanni Santia and Jake Williams. Buzzface: A news veracity dataset with facebook user commentary and egos. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1):531–540, Jun. 2018. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14985.

Sam Scott and Stan Matwin. Feature engineering for text classification. In *ICML*, volume 99, pages 379–388. Citeseer, 1999.

Clive Seale, Jonathan Charteris-Black, Aidan MacFarlane, and Ann McPherson. Interviews and internet forums: A comparison of two sources of qualitative data. *Qualitative Health Research*, 20(5):595–606, May 2010. ISSN 1049-7323. doi: 10.1177/1049732309354094.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1282.

O. Shahmirzadi, A. Lugowski, and K. Younge. Text similarity in vector space models: A comparative study. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 659–666, 2019. doi: 10.1109/ICMLA.2019.00120.

Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 745–750, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341448. doi: 10.1145/2872518.2890098. URL https://doi.org/10.1145/2872518.2890098.

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9(1):4787, Nov 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-06930-7.

K. Shu, S. Wang, and H. Liu. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435, 2018. doi: 10.1109/MIPR.2018.00092.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36, September 2017. ISSN 1931-0145. doi: 10.1145/3137597.3137600. URL http://doi.acm.org/10.1145/3137597.3137600.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188, 2020. doi: 10.1089/big.2020.0062. PMID: 32491943.

John Sinclair, Susan Jones, and Robert Daley. *English collocation studies: The OSTI report*. Bloomsbury Publishing, 2004. ISBN 9780826474896.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. Automatic detection of stance towards vaccination in online discussion forums. In *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pages 1–8, Taipei, Taiwan, November 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W17-5801.

Jonathan B Slapin and Sven-Oliver Proksch. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52, 2008. doi: https://doi.org/10.1111/j.1540-5907.2008.00338.x.

Robert R. Sokal and F. James Rohlf. The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40, 1962. ISSN 00400262. URL http://www.jstor.org/stable/1217208.

Jacob Soll. The long and brutal history of fake news, 2016. URL https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535/. accessed: 2021/12/18.

Alexander Spencer and Kai Oppermann. Narrative genres of Brexit: the Leave campaign and the success of romance. *Journal of European Public Policy*, 27, 2020. doi: https://doi.org/10.1080/13501763.2019.1662828.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-5107.

Bernd Carsten Stahl. On the difference or equality of information, misinformation, and disinformation: A critical research perspective. *Informing Science: The International Journal of an Emerging Transdiscipline*, 9:083–096, 2006. URL https://www.informingscience.org/Publications/473.

Constantina Stamou. Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Literary and Linguistic Computing*, 2007.

Ian Stewart and Jacob Eisenstein. Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4370, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1467.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1355.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012. URL https://www.isca-speech.org/archive/interspeech_2012/i12_0194.html.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

VIREN Swami, Jakob Pietschnig, ULRICH S. Tran, INGO W. Nader, Stefan Stieger, and Martin Voracek. Lunar lies: The impact of informational framing and individual differences in shaping conspiracist beliefs about the moon landings. *Applied Cognitive Psychology*, 27(1):71–80, 2013. doi: https://doi.org/10.1002/acp.2873.

Viren Swami, Martin Voracek, Stefan Stieger, Ulrich S. Tran, and Adrian Furnham. Analytic thinking reduces belief in conspiracy theories. *Cognition*, 133(3):572–585, 2014. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2014.08.006.

Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, August 2010. ISSN 0001-0782. doi: 10.1145/1787234.1787254.

Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. Some like it hoax: Automated fake news detection in social networks. *CoRR*, abs/1704.07506, 2017. URL http://arxiv.org/abs/1704.07506.

Chenhao Tan. Tracing community genealogy: How new communities emerge from the old. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, Jun 2018. URL https://ojs.aaai.org/index.php/ICWSM/article/view/15003.

Chenhao Tan and Lillian Lee. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 1056–1066, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741661. URL https://doi.org/10.1145/2736277.2741661.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1452.

Sheryl Thorburn and Laura M. Bogart. Conspiracy beliefs about birth control: Barriers to pregnancy prevention among african americans of reproductive age. *Health Education & Behavior*, 32(4):474–487, 2005. doi: 10.1177/1090198105276220.

James Thorne and Andreas Vlachos. An extensible framework for verification of numerical claims. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 37–40, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-3010.

James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1283.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-1074.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal, editors. *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, Brussels, Belgium, November 2018b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-5500.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal, editors. *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, Hong Kong, China, November 2019a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-6600.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. The FEVER2.0 shared task. In *Proceedings of the Second Workshop*

on Fact Extraction and VERification (FEVER), pages 1–6, Hong Kong, China, November 2019b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-6601.

Catalina L. Toma and Jeffrey T. Hancock. Reading between the lines: Linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 5–8, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-795-0. doi: 10.1145/1718918.1718921.

Leanne Townsend and Claire Wallace. Social media research: A guide to ethics. *University of Aberdeen*, 1:16, 2016. URL https://www.gla.ac.uk/media/Media_487729_smxx.pdf.

Trang Tran and Mari Ostendorf. Characterizing the language of online communities and its relation to community reception. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1030–1035, Austin, Texas, November 2016. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D16-1108.

Oren Tsur and Ari Rappoport. Don't let me be #misunderstood: Linguistically motivated algorithm for predicting the popularity of textual memes. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1), Apr. 2015. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14603.

Yulia Tsvetkov, Vinodkumar Prabhakaran, and Rob Voigt. Socially responsible NLP. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 24–26, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N18-6005.

William E Underwood, David Bamman, and Sabrina Lee. The transformation of gender in english-language fiction. *Journal of Cultural Analytics*, 1(1), 2 2018. doi: 10.22148/16.019.

Tanguy Urvoy, Emmanuel Chauveau, Pascal Filoche, and Thomas Lavergne. Tracking web spam with html style similarities. *ACM Transactions on the Web (TWEB)*, 2(1): 3, 2008.

Joseph E Uscinski. *Conspiracy theories and the people who believe them*. Oxford University Press, 2018. doi: 10.1093/oso/9780190844073.001.0001.

Reine C. van der Wal, Robbie M. Sutton, Jens Lange, and João P. N. Braga. Suspicious binds: Conspiracy thinking and tenuous perceptions of causal connections between co-occurring and spuriously correlated events. *European Journal of Social Psychology*, 48(7):970–989, 2018. doi: https://doi.org/10.1002/ejsp.2507.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. Monday mornings are my fave :) #not exploring the automatic recognition of irony in English tweets. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2730–2739, Osaka, Japan, December 2016a. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1257.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. Exploring the realization of irony in Twitter data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1794–1799, Portorož, Slovenia, May 2016b. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L16-1283.

Dirk van Hulle and Mike Kestemont. Periodizing samuel beckett's works: A stylochronometric approach. *Style*, 50(2):172–202, 2016. ISSN 00394238, 23746629. URL http://www.jstor.org/stable/10.5325/style.50.2.0172.

Trevor van Mierlo. The 1% rule in four digital health social networks: An observational study. *J Med Internet Res*, 16(2):e33, 2014. ISSN 14388871. URL http://www.jmir.org/2014/2/e33/.

Jan-Willem van Prooijen. Why education predicts decreased belief in conspiracy theories. *Applied Cognitive Psychology*, 31(1):50–58, 2017. doi: https://doi.org/10.1002/acp.3301.

Jan-Willem van Prooijen and Karen M. Douglas. Belief in conspiracy theories: Basic principles of an emerging research domain. *European Journal of Social Psychology*, 48(7):897–908, 2018. doi: https://doi.org/10.1002/ejsp.2530.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Alfredo Vellido Alcacena, Jose D Martin Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *ESANN 2012 proceedings: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning: Bruges (Belgium), 25-27 April, 2012*, pages 163–172, 2012.

Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA, June 2014. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W14-2508.

Andreas Vlachos and Sebastian Riedel. Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D15-1312.

Nguyen Vo and Kyumin Lee. The rise of guardians: Fact-checking url recommendation to combat fake news. In *The 41st International ACM SIGIR Conference on Research &; Development in Information Retrieval*, SIGIR '18, page 275–284, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210037.

Svitlana Volkova and Jin Yea Jang. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 575–583, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356404. doi: 10.1145/3184558.3188728.

Nigel Walker. Brexit timeline: events leading to the uk's exit from the european union. Technical report, House of Commons Library, 2021. URL https://commonslibrary.parliament.uk/research-briefings/cbp-7960/.

Byron C. Wallace. Computational irony: A survey and new perspectives. *Artificial Intelligence Review*, 43(4):467–483, Apr 2015. ISSN 1573-7462. doi: 10.1007/s10462-012-9392-5.

William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067.

Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 424–433, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933395. doi: 10.1145/1150402.1150450.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 849–857, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3219903.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. CrossWeigh: Training named entity tagger from imperfect annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5154–5163, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-1519.

Aleksander Wawer, Grzegorz Wojdyga, and Justyna Sarzyńska-Wawer. Fact checking or psycholinguistics: How to distinguish fake and true claims? In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 7–12, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-6602.

Nora Wenzl. 'this is about the kind of Britain we are': National identities as constructed in parliamentary debates about EU membership. In *Discourses of Brexit*, chapter 3. Routledge, 2019. doi: https://doi.org/10.4324/9781351041867.

Harry G. West and Todd Sanders. *Transparency and Conspiracy: Ethnographies of Suspicion in the New World Order*. Duke University Press, 03 2003. ISBN 978-0-8223-3036-3. doi: 10.1215/9780822384854.

Jennifer A. Whitson and Adam D. Galinsky. Lacking control increases illusory pattern perception. *Science*, 322(5898):115–117, 2008. ISSN 0036-8075. doi: 10.1126/science.1159845.

Richard J. Whitt, editor. *Diachronic Corpora, Genre, and Language Change*. John Benjamins, 2018. doi: 10.1075/scl.85.

Deirdre Wilson and Dan Sperber. On verbal irony. *Lingua*, 87(1):53–76, 1992.

Michael Wood and Karen Douglas. "what about building 7?" a social psychological study of online discussion of 9/11 conspiracy theories. *Frontiers in Psychology*, 4: 409, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00409.

Michael J. Wood, Karen M. Douglas, and Robbie M. Sutton. Dead and alive: Beliefs in contradictory conspiracy theories. *Social Psychological and Personality Science*, 3(6):767–773, Nov 2012. ISSN 1948-5506. doi: 10.1177/1948550611434786.

Simon N Wood. *Generalized additive models: an introduction with R*. Chapman and Hall, 2017. ISBN 9781498728331. URL `https://www.routledge.com/Generalized-Additive-Models-An-Introduction-with-R-Second-Edition/Wood/p/book/9781498728331`.

K. Wu, S. Yang, and K. Q. Zhu. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st International Conference on Data Engineering*, pages 651–662, 2015. doi: 10.1109/ICDE.2015.7113322.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, November 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-main.523`.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 1445–1456, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320351. doi: 10.1145/2488388.2488514.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5644–5651, 2019. doi: 10.1609/aaai.v33i01.33015644.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 673–681, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355810. doi: 10.1145/3159652.3159703.

Chunhui Yuan and Haitao Yang. Research on k-value selection method of k-means clustering algorithm. *J*, 2(2):226–235, 2019. ISSN 2571-8800. doi: 10.3390/j2020016.

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, page 405–417, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450351188. doi: 10.1145/3131365.3131390.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1007–1014, Republic and Canton of Geneva, CHE, 2018a. International World Wide Web Conferences Steering Committee. ISBN 9781450356404. doi: 10.1145/3184558.3191531.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, page 188–202, New York, NY, USA, 2018b. Association for Computing Machinery. ISBN 9781450356190. doi: 10.1145/3278532.3278550. URL https://doi.org/10.1145/3278532.3278550.

Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 218–226, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3316495.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. Community identity and user engagement in a multi-community landscape. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), May 2017. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14904.

Yi Zhang, Zachary Ives, and Dan Roth. "who said it, and why?" provenance for natural language claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4416–4426, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.406.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 1395–1405, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741637. URL https://doi.org/10.1145/2736277.2741637.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006. doi: https://doi.org/10.1002/asi.20316.

Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 2009.

Arkaitz Zubiaga and Heng Ji. Tweet, but verify: epistemic study of information verification on twitter. *Social Network Analysis and Mining*, 4(1):163, Mar 2014. ISSN 1869-5469. doi: 10.1007/s13278-014-0163-y.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. Towards detecting rumours in social media. In *AAAI Workshop: AI for Cities*, 2015. URL `https://arxiv.org/pdf/1504.04712.pdf`.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29, 03 2016. doi: 10.1371/journal.pone.0150989.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Exploiting context for rumour detection in social media. In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, editors, *Social Informatics*, pages 109–123, Cham, 2017. Springer International Publishing. ISBN 978-3-319-67217-5. URL `https://link.springer.com/chapter/10.1007/978-3-319-67217-5_8`.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Comput. Surv.*, 51(2), February 2018. ISSN 0360-0300. doi: 10.1145/3161603. URL `https://doi.org/10.1145/3161603`.