Proceedings of the 5th Workshop on

Making Sense of Microposts (#Microposts2015)

Big things come in small packages

at the 24th International Conference on the World Wide Web (WWW'15)
Florence, Italy
18th of May 2015

edited by
Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie

Preface

#Microposts2015, the 5th Workshop on *Making Sense of Microposts*, was held in Florence, Italy, on the 18th of May 2015, during (WWW'15), the 24th International Conference on the World Wide Web. The #Microposts journey started at the 8th Extended Semantic Web Conference (ESWC 2011, as #MSM, with the change in acronym from 2014), and moved to WWW in 2012, where it has stayed, for the fourth year now. *#Microposts2015* continues to highlight the importance of the medium, as we see end users appropriating Microposts, small chunks of information published online with minimal effort, as part of daily communication and to interact with increasingly wider networks and new publishing arenas.

The #Microposts workshops are unique in that they solicit participation not just from Computer Science, but encourage interdisciplinary work. We welcome research that looks at computational analysis of Microposts, as well as studies that employ mixed methods, and also those that examine the human generating and consuming Microposts and interacting with other users via this publishing venue. New to #Microposts2015 is a dedicated Social Sciences track, to encourage, particularly, contribution from the Social Sciences, to harness the advantages that approaches to analysing Microposts from this perspective bring to the field.

The term *Micropost* now rarely needs definition. Microposts are here to stay, and have evolved from text only, to include images, and now, audio and video. New platforms are developed each year to serve specific markets, and niche services compete with each other for a share of the audience. Twitter's *Periscope* is a new service similar to *Meerkat*, both of which use microblogging platforms to alert a network to a live video stream. Microposts now often serve also as a portal, and are harnessed by recommendation services, marketing and other enterprise to advertise or push information, products and services on other platforms. This is a not surprising means to access potential users, who now exchange Microposts round the clock, using a variety of publishing platforms. Media trends show that users are doing so increasingly from personal, mobile devices, as a preferred/convenient option that started to overtake usage on PCs in 2014. To extend reach, both in developed and emerging markets, services for publishing Microposts from feature phones are being developed – these include the usual suspects, Twitter and Facebook, who employ native apps or the mobile web, and also newer entrants with dedicated services and apps such as *Saya*. Country and language-specific platforms such as Sina Weibo, while not as widespread, serve a specific region and market, especially where any of a number of reasons prevent access to the more well-known microblogging platforms. Political movements such as the Arab Spring have been reported to have increased the use of social media services and microblogging particularly in regions concerned, as the quick, low-cost means for sharing, in the moment, breaking news, local and context-specific information and personal stories, resulted in an increased sense of community and solidarity. Interestingly, in response to emergencies, mass demonstrations and other social events such as festivals and conferences, when regular access to communication services is often interrupted and/or unreliable, developers are quick to offer alternatives that end users piggyback on to post information. *Line* was born to serve such a need, to provide an alternative communication service and support emergency response during a natural disaster in Japan in 2011. Its popularity continued beyond its initial purpose, and Line has grown into a popular (regional) microblogging service.

The #Microposts workshop was created to bring together researchers in different fields studying the publication, analysis and reuse of these very small chunks of information, shared in private, semi-public and fully open, social and formal networks. Microposts collectively make up a vast knowledge store, contained in what is today described as "big data" – heterogenous, increasing at phenomenal rates, and with multiple, unbridled authors, covering myriad topics with varying degrees of accuracy and veracity. With each year we have seen submissions tackling different aspects of Microposts, with new methods and techniques developed to analyse this valuable dataset and also its publishers, human or bot, and examining the different ways in which the medium is used. With the increase in the use of Microposts as a portal to other services, we saw, this year, studies on the detection and analysis of spam, and the use of open posting as a cover for disseminating extremist opinions or to swamp dissenting views. Reflecting the very social nature of the publishing platform, submissions also covered analysis of the human reaction to recent, provoking news events.

We thank all contributors and participants: each author's work adds to research that continues to advance the field. Submissions to the two research tracks came from institutions in ten countries around the world. The challenge also continues to see wide interest, with final submissions from academia and industry, across six countries. Our programme committee is even more varied, working in academia, independent research institutions and industry, and spanning an even larger number of countries. Most of our PC have reviewed for more than one, and a good percentage, all five #Microposts workshops. Very special thanks to our committee, without whom we would not be able to run the workshop – their dedication is seen in the feedback provided to us and to authors. Thanks also to the chairs of the Social Sciences Track and the NEEL Challenge, whose work has been invaluable in pulling the three parts together into a unified, successful workshop.

Matthew Rowe Lancaster University, UK
Milan Stankovic Sépage / Université Paris-Sorbonne, France
Aba-Sah Dadzie KMi, The Open University, UK
#Microposts2015 Organising Committee, May 2015

Introduction to the Proceedings

Main Track

The main workshop track attracted nine submissions, out of which two long papers and one short were accepted, in addition to an extended abstract and a poster. It should be noted that two of these crossed the boundary between Computer and Social Sciences, and were therefore assigned reviewers from both tracks. Topics covered ranged from machine learning and named entity recognition to Micropost classification and extraction. Applications were seen in topic, event and spam detection. We provide a brief introduction to each below.

De Boom, Van Canneyt & Dhoedt, in *Semantics-driven Event Clustering in Twitter Feeds*, present a novel perspective on event-detection in tweets, by associating semantics to tweets and hashtags. They demonstrate how an approach that combines machine learning with explicit semantics detection can yield considerable improvement over state of the art event clustering approaches.

In the paper *Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter*, Wang, Zubiaga, Liakata & Procter investigate the use of a variety of feature sets and classifiers for the detection of social spam on Twitter. These include user features (social network properties of the tweeter, such as their in- and out-degrees); content features (number of hashtags and mentions); n-gram features (mined from textual aspects); and sentiment features, based on both manually and automatically created semantic lexicons. Classifiers tested including naïve Bayes; k-Nearest Neighbours, Support Vector Machines, Decision Trees, and Random Forests. The paper presents an interesting investigation, classifying users as spammers (or not), as opposed to existing work which attempts to classify content as spam (or not).

In *User Interest Modeling in Twitter with Named Entity Recognition*, Karatay & Karagoz explore techniques for user profiling using Named Entity detection in tweets – a topic of increasing importance in the era of information overload, where filtering and personalising information is crucial for user engagement and experience. The in-depth view of appropriate techniques and issues related to Named Entity-based user profiling on Twitter will interest both academic and industrial audiences.

Within the broader area of spam, misconduct and automated accounts on Twitter, Edwards & Guy study the *Connections between Twitter Spammer Categories*. Unlike most other work in this area, they do not only distinguish spam from non-spam, but assume there are different types of spam accounts, which they categorise as "advertising", "explicit", "follower gain", "celebrity" and "bot". They show, in their extended abstract, that each type of spammer behaves differently with respect to establishing follower relations with other spam accounts. They also observe that genuine Twitter users can be found as followers of all types of spam accounts, but are more likely to connect with specific types of spammers.

Agarwal & Sureka, in *A Topical Crawler for Uncovering Hidden Communities of Extremist Micro-Bloggers on Tumblr*, discuss the use of microblogging systems such as Tumblr to promote extremism, taking advantage of the ability to post information anonymously. The poster paper describes a process that uses pre-identified keywords to flag relevant posts, and hence, identify suspect tags in textual posts. A random walk from a seed blogger is then used to identify further individuals and communities promoting extremism. The authors report misclassification of 13% and accuracy of 77% for predicting "hate promoting bloggers", with misclassification of unknown bloggers at 34%.

Social Sciences Track

The Social Sciences track attracted three submissions, of which two were accepted. In addition to data mining and/or statistical analysis over the very large amounts of data involved, each submission carried out in-depth, qualitative analysis to tease out nuanced information that is more difficult to identify with automated methods. The track was chaired by Katrin Weller and Danica Radovanović.

One of the major contemporary events that spiked user engagement on social media during the first months of 2015 was the Charlie Hebdo shooting in France on January 7th. Giglietto & Lee provide one of the first studies of Twitter users' reactions to this event, in *To Be or Not to Be Charlie: Twitter Hashtags as a Discourse and Counter-discourse in the Aftermath of the 2015 Charlie Hebdo Shooting in France*. In particular, they study the use of the hashtag #JeNeSuisPasCharlie, which was used in contrast to the initial #JeSuisCharlie hashtag. Using different approaches to data analysis (including activity patterns and word frequencies) the authors demonstrate how tweets including #JeNeSuisPasCharlie rather resemble crisis communication patterns, and at the same time support different expressions of self-identity such as grief and resistance.

Coelho, Lapa, Ramos & Malini, in *A Research Design for the Analysis of Contemporary Social Movements*, present a research method to identify elements that promote social empowerment in the political vitality present in digital culture. They developed a model of investigation that allows discursive analysis of posts generated within net activist groups. Methods, instruments and resources were created and articulated for the collection and treatment of big data and for further qualitative analysis of content. In addition to contributing to ICT, by proposing a qualitative investigation of social networks, this research design contributes to the field of Education, as the results of its application can be used be used to develop guidelines for teachers, to support critical appropriation and education of social networks.

Named Entity rEcognition & Linking (NEEL) Challenge

The #Microposts2015 NEEL challenge again increased in complexity, to address further challenges encountered in the analysis of Micropost data. This year's challenge required participants to recognise entities and their types, and also link them, where found, to corresponding DBpedia resources.

The challenge attracted good interest from the community, with 29 intents to submit, out of which 21 applied for the final evaluation. Seven took part in the quantitative evaluation and six completed submission (including a written abstract). Of these three were accepted for presentation and a further three as posters. All accepted submissions also took part in the workshop's poster session, whose aim is to exhibit practical application in the field and foster further discussion about the ways in which knowledge content is extracted from Microposts and reused.

The NEEL challenge was chaired by A. Elizabeth Cano and Giuseppe Rizzo, with Andrea Varga and Bianca Pereira as dataset chairs. As in previous years, the challenge committee prepared a gold standard from the challenge corpus, which covered events in 2011, '13 & 14 on, for example, the London Riots, the Oslo bombing and the UCI Cyclo-cross World Cup. Changes to the submission and evaluation protocols included wrapping submissions as a publicly accessible, REST-based service. Up to ten runs were allowed per submission, of which the best three were used in computing the final rankings, using four weighted metrics: tagging (0.3), linking (0.3), clustering (0.4) and latency (computation time) to sort in case of a tie.

We provide here a brief introduction to participants' abstracts describing their submissions, and more detail about the preparation and evaluation processes in the challenge summary paper included in the proceedings.

Yamada, Takeda & Takefuji, in *An End-to-End Entity Linking Approach for Tweets*, present a five stage approach: (1) preprocessing, (2) candidate mention generation, (3) mention detection and disambiguation, (4) NIL mention detection and (5) type prediction. In preprocessing, they utilise tokenisation and POS tagging based on state of the art algorithms, along with extraction of tweet timestamps. Yamada *et al.* tackle candidate mention generation and disambiguation using fuzzy search of Wikipedia for candidate entity mentions, and popularity of Wikipedia pages for ranking the set of candidate entities. Finally, they tackle selection of NIL mentions and entity typing as supervised learning problems.

In *Entity Recognition and Linking on Tweets with Random Walks*, Guo & Barbosa present a sequential approach to the NEEL task by, first, recognising entities using *TwitIE*, and then linking them to corresponding DBpedia entities. Starting from the (DBpedia) candidate entities, Guo & Barbosa build a subgraph by adding all adjacent entities to the candidates. They execute a personalised PageRank, giving more importance to unambiguous entities. They then measure semantic relatedness between entity candidates and the "unambiguous" entities for the "document", and employ threshold and name similarity for NIL prediction and clustering.

In the submission *Combining Multiple Signals for Semanticizing Tweets: University of Amsterdam at #Microposts2015*, Gârbacea, Odijk, Graus, Sijaranamual & de Rijke employ a sequential approach composed of four stages: (1) candidate mention detection, (2) candidate typing and linking, (3) NIL clustering and (4) overlap resolution. The first stage is tackled with an annotation-based process that takes as input the lexical content of Wikipedia and an NER classifier trained using the challenge dataset. To resolve candidate mention overlaps, the authors propose an algorithm based on the results of the linking stage and the *Viterbi* path resolution output. A "learning to rank" supervised model is used to select the most representative DBpedia reference entity, and, therefore, type of each candidate mention, normalising the type via manual alignment from the DBpedia ontology and the NEEL taxonomy. Finally, Gârbacea *et al.* solve the NIL using a clustering algorithm operating on the lexical similarity of the candidate mentions for which no counterparts are found in DBpedia.

Basile, Caputo & Semeraro in *UNIBA: Exploiting a Distributional Semantic Model for Disambiguating and Linking Entities in Tweets*, introduce an unsupervised approach which uses a modified version of their *Lesk* algorithm. Basile *et al.* use similarity of "distributional semantic spaces" for disambiguation, and two alternative and state of the art approaches for the candidate identification phase, based on either POS tagging or n-gram similarity. Entities are typed through inheritance of the type of the DBpedia reference entity pointed to, which is in turn manually aligned to the NEEL taxonomy.

In *AMRITA - CEN@NEEL: Identification and Linking of Twitter Entities*, Barathi Ganesh, Abinaya, Anand Kumar, Soman & Vinaykumar address the NEEL task sequentially by, first, tokenising and tagging the tweets using *TwitIE*. They then classify entity mentions by applying supervised learning using direct (POS tags) and indirect features (the two words before and after a candidate mention entity). Using a total of 34 lexical features, the authors experiment with three supervised learning algorithms to determine the recognition configuration that would achieve the best performance in the development test. Barathi Ganesh *et al.* tackle the linking task by looking up DBpedia reference entries; that maximising the similarity score between related entries and the named entities is designated the representative. Named entities without related links are assigned as NIL.

Finally, Sinha & Barik, in *Named Entity Extraction and Linking in #Microposts*, present a sequential approach to the NEEL task which recognises entities and then links them. The first stage is grounded on linguistic clues extracted from conventional approaches such as POS tagging, word capitalisation and hashtag in the tweet. They then train a CRF with the linguistic features and the contextual similarity of adjacent tokens, with the token window set to 5. Priyanka & Barik perform the linking task using an entity resolution mechanism that takes as input the output of the NER stage and that of *DBpedia Spotlight*. For each entity returned from DBpedia Spotlight found to be a substring of any of the entities extracted in the NER stage and for which a substring match is found, the corresponding URI is returned and assigned to it. Otherwise the entity is assigned as NIL.

Workshop Awards

Main Track. The #Microposts2015 best paper award went to:

Cedric De Boom, Steven Van Canneyt & Bart Dhoedt
for their submission entitled:

Semantics-driven Event Clustering in Twitter Feeds

Social Sciences Track. GESIS[1], the Leibniz Institute for the Social Sciences, sponsored the best paper award for the Social Sciences track. We teamed up with GESIS, the largest service and infrastructure institution for the Social Sciences in Germany, to highlight the role of interdisciplinary approaches in obtaining a better understanding of the users behind social media and Microposts. As in the main track, the decision was guided by nominations from the reviewers and review scores. The #Microposts2015 Social Sciences Track best paper award went to:

Fabio Giglietto & Yenn Lee
for their submission entitled:

To Be or Not to Be Charlie: Twitter Hashtags as a Discourse and Counter-discourse in the Aftermath of the 2015 Charlie Hebdo Shooting in France

NEEL Challenge. SpazioDati[2], an Italian startup who took part in the #Microposts2014 NEEL challenge, sponsored the award for the best submission. SpazioDati aim to provide access to a single source of common-sense knowledge, mined and synthesised from a large number of open and closed data sources. By sponsoring the challenge, SpazioDati reinforce the value in the content of the increasingly large knowledge source that is Micropost data. The challenge award was also determined by the results of the quantitative evaluation. The #Microposts NEEL Challenge award went to:

Ikuya Yamada, Hideaki Takeda & Yoshiyasu Takefuji
for their submission entitled:

An End-to-End Entity Linking Approach for Tweets

[1] http://www.gesis.org
[2] http://spaziodati.eu

Additional Material

The call for participation and all paper, poster and challenge abstracts are available on the #Microposts2015 website[3]. The full proceedings are also available on the CEUR-WS server, as Vol-1395[4]. The gold standard for the NEEL Challenge is available for download[5].

The proceedings for #Microposts2014 are available as Vol-1141[6] The proceedings for the #MSM2013 main track are available as part of the WWW'13 Proceedings Companion[7]. The #MSM2013 Concept Extraction Challenge proceedings are published as a separate volume as CEUR Vol-1019[8], and the gold standard is available for download[9]. The proceedings for #MSM2012 and #MSM2011 are available as CEUR Vol-838[10]. and CEUR Vol-718[11], respectively.

#Microposts2015

gesis Leibniz Institute for the Social Sciences

SPAZIODATI

[3] http://www.scc.lancs.ac.uk/microposts2015
[4] #Microposts2015 Proc. http://ceur-ws.org/Vol-1395
[5] http://ceur-ws.org/Vol-1395/microposts2015_neel-challenge-report/microposts2015-neel_challenge_gs.zip
[6] #Microposts2014 Proc. http://ceur-ws.org/Vol-1141
[7] WWW'13 Companion: http://dl.acm.org/citation.cfm?id=2487788
[8] #MSM2013 CE Challenge Proc. http://ceur-ws.org/Vol-1019
[9] http://ceur-ws.org/Vol-1019/msm2013-ce_challenge_gs.zip
[10] #MSM2012 Proc. http://ceur-ws.org/Vol-838
[11] #MSM2011 Proc. http://ceur-ws.org/Vol-718

Main Track Programme Committee

Pierpaolo Basile University of Bari, Italy
Julie Birkholz CHEGG, Ghent University, Belgium
John Breslin National University of Ireland Galway, Ireland
A. Elizabeth Cano KMi, The Open University, UK
Marco A. Casanova Pontifícia Universidade Católica do Rio de Janeiro, Brazil
Óscar Corcho Universidad Politécnica de Madrid, Spain
Guillaume Erétéo Vigiglobe, France
Miriam Fernandez KMi, The Open University, UK
Andrés Garcia-Silva Universidad Politécnica de Madrid, Spain
Anna Lisa Gentile The University of Sheffield, UK
Jelena Jovanovic University of Belgrade, Serbia
Mathieu Lacage Alcméon, France
Philipe Laublet Université Paris-Sorbonne, France
José M. Morales del Castillo El Colegio de México, Mexico
Fabrizio Orlandi University of Bonn, Germany
Bernardo Pereira Nunes Pontifícia Universidade Católica do Rio de Janeiro, Brazil
Danica Radovanović University of Belgrade, Serbia
Guiseppe Rizzo Eurecom, France
Harald Sack HPI, University of Potsdam, Germany
Bernhard Schandl mySugr GmbH, Austria
Sean W. M. Siqueira Universidade Federal do Estado do Rio de Janeiro, Brazil
Victoria Uren Aston Business School, UK
Andrea Varga Swiss Re, UK
Katrin Weller GESIS Leibniz Institute for the Social Sciences, Germany
Alistair Willis The Open University, UK
Ziqi Zhang The University of Sheffield, UK

Sub Reviewers

Tamara Bobic HPI, University of Potsdam, Germany

Social Sciences Track Programme Committee

Gholam R. Amin Sultan Qaboos University, Oman
Julie Birkholz CHEGG, Ghent University, Belgium
Tim Davies University of Southampton, UK
Munmun De Choudhury Georgia Tech, USA
Ali Emrouznejad Aston Business School, UK
Fabio Giglietto Università di Urbino Carlo Bo, Italy
Simon Hegelich Universität Siegen, Germany
Kim Holmberg University of Turku, Finland
Athina Karatzogianni University of Leicester, UK
José M. Morales del Castillo El Colegio de México, Mexico
Raquel Recuero Universidade Católica de Pelotas, Brazil
Bianca C. Reisdorf University of Leicester, UK
Luca Rossi Università di Urbino Carlo Bo, Italy
Saskia Vanmanen The Open University, UK
Alistair Willis The Open University, UK
Taha Yasseri University of Oxford, UK
Victoria Uren Aston Business School, UK

NEEL Challenge Evaluation Committee

Gabriele Antonelli SpazioDati, Italy
Ebrahim Bagheri Ryerson University, Canada
Pierpaolo Basile University of Bari, Italy
Grégoire Burel KMi, The Open University, UK
Óscar Corcho Universidad Politécnica de Madrid, Spain
Leon Derczynski The University of Sheffield, UK
Milan Dojchinovski Czech Technical University in Prague, Czech Republic
Guillaume Erétéo Vigiglobe, France
Andrés Garcia-Silva Universidad Politécnica de Madrid, Spain
Anna Lisa Gentile The University of Sheffield, UK
Miguel Martinez-Alvarez Signal, UK
José M. Morales del Castillo El Colegio de México, Mexico
Bernardo Pereira Nunes Pontifícia Universidade Católica do Rio de Janeiro, Brazil
Daniel Preoţiuc-Pietro University of Pennsylvania, USA
Giles Reger Otus Labs, UK
Irina Temnikova Qatar Computing Research Institute, Qatar
Victoria Uren Aston Business School, UK

Table of Contents

Section IVa: NEEL Challenge Submissions I

Section IVb: NEEL Challenge Submissions II – Posters

Section I:

Main Research Track

Semantics-driven Event Clustering in Twitter Feeds

Cedric De Boom
Ghent University – iMinds
Gaston Crommenlaan 8-201,
9050 Ghent, Belgium
cedric.deboom@intec.ugent.be

Steven Van Canneyt
Ghent University – iMinds
Gaston Crommenlaan 8-201,
9050 Ghent, Belgium
steven.vancanneyt@intec.ugent.be

Bart Dhoedt
Ghent University – iMinds
Gaston Crommenlaan 8-201,
9050 Ghent, Belgium
bart.dhoedt@intec.ugent.be

ABSTRACT

Detecting events using social media such as Twitter has many useful applications in real-life situations. Many algorithms which all use different information sources—either textual, temporal, geographic or community features—have been developed to achieve this task. Semantic information is often added at the end of the event detection to classify events into semantic topics. But semantic information can also be used to drive the actual event detection, which is less covered by academic research. We therefore supplemented an existing baseline event clustering algorithm with semantic information about the tweets in order to improve its performance. This paper lays out the details of the semantics-driven event clustering algorithms developed, discusses a novel method to aid in the creation of a ground truth for event detection purposes, and analyses how well the algorithms improve over baseline. We find that assigning semantic information to every individual tweet results in just a worse performance in F_1 measure compared to baseline. If however semantics are assigned on a coarser, hashtag level the improvement over baseline is substantial and significant in both precision and recall.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Experimentation

Keywords

Semantic information, event detection, clustering, social media, Twitter

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

1. INTRODUCTION

Traditional media mainly cover large, general events and thereby aim at a vast audience. Events that are only interesting for a minority of people are rarely reported. Next to the traditional mass media, social media such as Twitter and Facebook are a popular source of information as well, but extracting valuable and structured data from these media can be challenging. Posts on Twitter for example have a rather noisy character: written text is mostly in colloquial speech full of spelling errors and creative language use, such posts often reflect personal opinions rather than giving an objective view of the facts, and a single tweet is too short to grasp all the properties that represent an event. Nevertheless the user-contributed content on social media is extensive, and leveraging this content to detect events can complement the news coverage by traditional media, address more selective or local audiences and improve the results of search engines.

In the past researchers mostly used textual features as their main source of information to perform event detection tasks in social media posts. Next to the text itself, other characteristic features such as the timestamp of the post, user behavioural patterns and geolocation have been successfully taken into account [1, 4, 15, 17, 18, 22]. Less used are so-called semantic features, in which higher-level categories or semantic topics are captured for every tweet and used as input for the clustering algorithm. These semantic topics can either be very specific—such as sports, politics, disasters...—or can be latent abstract categories not known beforehand; such an abstract topic is usually a collection of semantically related words. In most applications semantics are determined on event level after the actual event detection process [19]. We however propose to use semantic information on tweet level to drive the event detection algorithm. After all, events belonging to different semantic categories—and thus also its associated tweets—are likely to be discerned more easily than semantically related events. For example then it is relatively easy to distinguish the tweets of a sports game and a concurrent politics debate.

The use case we address in this paper consists of dividing a collection of tweets into separate events. In this collection every tweet belongs to a certain event and it is our task to cluster all tweets in such a way that the underlying event

structure is reflected through these clusters of tweets. For this purpose we adopt a single pass clustering mechanism. As a baseline we use a clustering approach which closely resembles the algorithm proposed by Becker et al. to cluster Flickr photo collections into events [2, 3], and in which we only use plain textual features. We then augment this baseline algorithm, now incorporating semantic information about the tweets as a second feature next to the text of the tweet. As it turns out, solely using a semantic topic per tweet only marginally improves baseline performance; the attribution of semantic labels on tweet level seems to be too fine-grained to be of any predictive value. We therefore employ an online dynamic algorithm to assign semantic topics on hashtag level instead of tweet level, which results in a courser attribution of topic labels. As will be shown in this paper, the latter approach turns out to be significantly better than baseline performance.

The remainder of this paper is structured as follows. In Section 2 we shortly discuss the most appropriate related work in recent literature, after which we describe the methodology to extract events from a collection of Twitter posts in Section 3. The collection of data and the construction of a ground truth is treated in Section 4. Finally we analyse the results of the developed algorithms in Section 5.

2. RELATED WORK

Since the emergence of large-scale social networks such as Twitter and their growing user base, the detection of events using social information has attracted the attention of the scientific community. In a first category of techniques, Twitter posts are clustered using similarity measures. These can be either based on textual, temporal, geographical or other features. Becker et al. were among the first to implement this idea by clustering a Flickr photo collection [2, 3]. They developed a single pass unsupervised clustering mechanism in which every cluster represented a single event. Their approach however scaled exponentially in the number of detected events, leading to Reuter et al. improving their algorithm by using a prior candidate retrieval step [15], thereby reducing the execution time to linear scaling. Petrović et al. used a different technique based on Locality Sensitive Hashing, which can also be seen as a clustering mechanism [14]. In this work, tweets are clustered into buckets by means of a hashing function. Related tweets are more probable to fall into the same bucket, which allows for a rapid comparison between tweets to drive the event detection process.

The techniques in a second category of event detection algorithms mainly use temporal and volumetric information about the tweets being sent. Yin et al. for example use a peak detection strategy in the volume of tweets to detect fire outbreaks [22], and Nichols et al. detect volume spikes to identify events in sporting games [13]. By analysing communication patterns between Twitter users, such as peaks in original tweets, retweets and replies, Chierichetti et al. were able to extract the major events from a World Cup football game or the Academy Awards ceremony [7]. Sakaki et al. regarded tweets as individual sensor points to detect earthquakes in Japan [17]. They used a temporal model to detect spikes in tweet volume to identify individual events, after which a spatial tracking model, such as a Kalman filter or a particle filter, was applied to follow the earthquake events as they advanced through the country. Bursts of words in time or in geographic location can also be calculated by using signal processing techniques, e.g. a wavelet transformation. Such a technique was successfully used by Weng et al. in their EDCoW algorithm to detect Twitter events [21], and by Chen and Roy to detect events in Flickr photo collections on a geographic scale [6].

Semantic information is often extracted after the events are detected to classify them into high level categories [16]. This can be done in either a supervised way, using a classifier like Naive Bayes or a Support Vector Machine, but most of the times unsupervised methods are preferred, since they do not require labelled data to train models and are able to discover semantic categories without having to specify these categories beforehand. Popular unsupervised techniques are Latent Dirichlet Allocation (LDA), clustering, Principal Component Analysis (PCA) or a neural auto-encoder. LDA was introduced by Blei et al. in 2003 as a generative model to extract latent topics from a large collection of documents [5]. Since then many variants of LDA have emerged tailored to specific contexts. Zhao et al. created the TwitterLDA algorithm to extract topics from microposts, such as tweets, assuming a tweet can only have one topic. Using community information next to purely textual information, Liu et al. developed their own version of LDA as well, called Topic-LinkLDA [10]. A temporal version of LDA, called TM-LDA, was developed by Wang et al. to be able to extract topics from text streams, such as a Twitter feed [20]. By batch grouping tweets in hashtag pools, Mehrotra et al. were able to improve standard LDA topic assignments to individual tweets [12].

3. EVENT CLUSTERING

In this section we will describe the mechanics to discover events in a collection of tweets. In the dataset we use, every tweet t is assigned a set of event labels E_t. This set contains more than one event label if the tweet belongs to multiple events. The dataset itself consists of a training set T_{train} and a test set T_{test}. The details on the construction of the dataset are found in Section 4. We will now try to recover the events in the test set by adopting a clustering approach. First the mechanisms of an existing baseline algorithm will be expounded. Next we will extend this algorithm using semantic information calculated from the tweets.

3.1 Baseline: Single Pass Clustering

Our baseline algorithm will use single pass clustering to extract events from the dataset. Becker et al. elaborated such an algorithm to identify events in Flickr photo collections [2, 3]; their approach was criticized and improved by Reuter et al. for the algorithm to function on larger datasets [15]. In this paper we will adopt single-pass clustering as a baseline that closely resembles the algorithm used by Becker et al.

As a preprocessing step, every tweet in the dataset is represented by a plain tf-idf vector and sorted based on its timestamp value. In the following we will use the same symbol t for the tweet itself and for its tf-idf vector. As the algorithm proceeds, it will create clusters of tweets, which are the retrieved events. We denote the cluster to which tweet t belongs as S_t; this cluster is also characterized by a cluster center point s_t. We refer to a general cluster and corre-

sponding cluster center point as resp. S and s. The set A contains all clusters which are currently active, i.e. being considered in the clustering procedure. During execution of the algorithm, a cluster is added to A if it is newly created. After some time a cluster can become inactive by removing this cluster from the set A. In Section 5 we will specify how a cluster can become inactive.

The baseline algorithm works as follows. When the current tweet t is processed, the cosine similarity $\cos(t, s)$ between t and cluster center s is calculated for all S in A. A candidate cluster S'_t (with cluster center s'_t) to which t could be added, and the corresponding cosine similarity $\cos(t, s'_t)$, are then calculated as

$$S'_t = \arg\max_{S \in A} \cos(t, s), \tag{1}$$

$$\cos(t, s'_t) = \max_{S \in A} \cos(t, s). \tag{2}$$

If S'_t does not exist—this occurs when A is empty—we assign t to a new empty cluster S_t, we set $s_t = t$ and S_t is added to A. If S'_t does exist, we need to decide whether t belongs to this candidate cluster or not. For this purpose we train a logistic regression classifier from LIBLINEAR [8] with a binary output. It takes $\cos(s'_t, t)$ as a single feature and decides whether t belongs to S'_t. If it does, then we set S_t to S'_t and we update its cluster center s_t as follows:

$$s_t = \frac{\sum_{t \in S_t} t}{|S_t|}. \tag{3}$$

If t does not belong to S'_t according to the classifier, then as before we assign t to a new empty cluster S_t and we set $s_t = t$.

In the train routine we assign every tweet one by one to a cluster corresponding to their event label. At every step we calculate the candidate cluster S'_t for every tweet t in T_{train} and verify whether this cluster corresponds to one of the event labels of t in the ground truth. If it does, we have a positive train example, otherwise a negative example. The number of positive and negative examples are balanced by randomly removing examples from either the positive or negative set, after which the examples are used to train the classifier.

In the original implementation by Becker et al. the processing of a tweet is far from efficient since every event cluster has to be tested. After a certain time period, the amount of clusters becomes very large. The adjustments by Reuter et al. chiefly aim at improving this efficiency issue. We do not consider these improvements here, since in Equation (1) we only test currently active clusters, which is already a performance gain.

3.2 Semantics-driven Clustering

To improve the baseline single pass clustering algorithm we propose a clustering algorithm driven by the semantics of the tweets. For example tweets that belong to the same semantic topic—e.g. sports, disasters, ...—are more likely to belong to the same event than tweets about different topics. Discerning two events can become easier as well if the two events belong to different categories.

To calculate a semantic topic for each of the tweets in the dataset, we make use of the TwitterLDA algorithm [23]. It is an adjustment of the original LDA (Latent Dirichlet Allocation) algorithm [5] for short documents such as tweets, in which every tweet only gets assigned a single topic—instead of a probabilistic distribution over all the topics—and single user topic models are taken into account. After running the TwitterLDA algorithm, every tweet t gets assigned a semantic topic γ_t.

The actual clustering algorithm has the same structure as the baseline algorithm, but it uses the semantic topic of the tweets as an extra semantic feature during clustering. We define the semantic fraction $\sigma(t, S)$ between a tweet and an event cluster as the fraction of tweets in S that have the same semantic topic as t:

$$\sigma(t, S) = \frac{|\{t' : t' \in S \wedge \gamma_{t'} = \gamma_t\}|}{|S|}. \tag{4}$$

To select a candidate cluster S'_t (with cluster center s'_t) to which t can be added, we use the cosine similarity, as before, as well as this semantic fraction:

$$S'_t = \arg\max_{S \in A} \cos(t, s) \cdot \sigma(t, S). \tag{5}$$

We choose to multiply cosine similarity and semantic fraction to select a candidate cluster since both have to be as large as possible, and if one of the two factors provides serious evidence against the candidate cluster, we want this to be reflected. Now we use both $\cos(t, s'_t)$ and $\sigma(t, S'_t)$ features to train a logistic regression classifier with a binary output. The rest of the algorithm continues in the way the baseline algorithm does.

3.3 Hashtag-level Semantics

As pointed out by Mehrotra et al. the quality of topic models on Twitter data can be improved by assigning topics to tweets on hashtag level instead of on tweet level [12]. To further improve the semantics-driven clustering, we therefore use a semantic majority voting scheme on hashtag level, which differs from the approach by Mehrotra et al. in that it can be used in an online fashion and that we consider multiple semantic topics per tweet.

In the training set we assign the same topic to all tweets sharing the same event label by performing a majority vote:

$$\forall t \in T_{\text{train}} : \gamma_t = \arg\max_{\gamma} \left| \left\{ t' : \gamma_{t'} = \gamma \wedge E_{t'} \cap E_t \neq \emptyset \right\} \right|. \tag{6}$$

This way every tweet in the training set is represented by a semantic topic that is dominated on the level of the events instead of on tweet level, resulting in a much coarser attribution of semantic labels. We cannot do this for the test set, since we do not know the event labels for the test set while executing the algorithm. We can however try to emulate such a majority voting at runtime. For this purpose, every tweet t is associated with a set of semantic topics Γ_t. We initialize this set as follows:

$$\forall t \in T_{\text{test}} : \Gamma_t = \{\gamma_t\}. \tag{7}$$

Next to a set of topics for every tweet, we consider a dedicated hashtag pool H_h for every hashtag h, by analogy with

[12]. With every pool H we associate a single semantic topic β_H. As the algorithm proceeds, more and more hashtag pools will be created and filled with tweets.

When a tweet t is processed in the clustering algorithm, it will first be added to some hashtag pools, depending on the number of hashtags in t. So for every hashtag h in t, t is added to H_h. When a tweet t is added to a hashtag pool H, a majority vote inside this pool is performed:

$$\beta_{\text{new},H} = \arg\max_{\gamma} \left|\left\{t' : t' \in H \wedge \gamma_{t'} = \gamma\right\}\right|. \tag{8}$$

We then update Γ_t for every tweet t in H:

$$\forall t \in H : \Gamma_{\text{new},t} = (\Gamma_{\text{old},t} \setminus \{\beta_H\}) \cup \{\beta_{\text{new},H}\}. \tag{9}$$

Finally $\beta_{\text{new},H}$ becomes the new semantic topic of H. Note that every tweet t keeps its original semantic topic γ_t.

What still needs adjustment in order for the clustering algorithm to use this new information, is the definition of the semantic fraction from Equation (4). We altered the definition as follows:

$$\sigma'(t, S) = \max_{g \in \Gamma_t} \frac{|\{t' : t' \in S \wedge g \in \Gamma_{t'}\}|}{|S|}. \tag{10}$$

Since Equation (10) implies Equation (4) if Γ_t contains only one element for every tweet t, this is a justifiable generalization.

4. DATA COLLECTION AND PROCESSING

In the past many datasets have been assembled to perform event clustering on social media. Unfortunately many of these datasets are not publicly available; this is especially true for Twitter datasets. We therefore choose to build our own dataset, available at `http://users.ugent.be/~cdboom/events/dataset.txt`. To speed up this task we follow a semi-manual approach, in which we first collect candidate events based on a hashtag clustering procedure, after which we manually verify which of these correspond to real-world events.

4.1 Event Definition

To identify events in a dataset consisting of thousands of tweets, we state the following event definition, which consists of three assumptions. Assumption 1 – a real-world event is characterized by one or multiple hashtags. For example, tweets on the past FIFA world cup football matches were often accompanied by hashtags such as #USAvsBelgium and #WorldCup. Assumption 2 – the timespan of an event cannot transgress the boundaries of a day. This means that if a certain real-world event takes place at several days—such as a music festival—this real-world event will be represented by multiple event labels. The assumption will allow us to discern events that share the same hashtag, but occur on a different day of the week, and will speed up the eventual event detection process. The hashtag #GoT for example will spike in volume whenever a new episode of Game of Thrones is aired, which are thus different events according to our definition. Assumption 3 – there is only one event that corresponds to a certain hashtag on a given day.

Assumption 3 is not restrictive and can easily be relaxed. For example if we would relax this Assumption and allow multiple events with the same hashtags to happen on the same day, we would need a feature in the event detection process to incorporate time differences, which is easily done. Alternatively we could represent our tweets using df-idf$_t$ vectors, instead of tf-idf vectors, which also consider time aspects of the tweets [1].

4.2 Collecting Data

We assembled a dataset by querying the Twitter Streaming API for two weeks, between September 29 and October 13 of the year 2014. We used a geolocation query and required that the tweets originated from within the Flanders region in Belgium, at least by approximation. Since only very few tweets are geotagged, our dataset was far from a representative sample of the tweets sent during this fortnight.

We therefore augment our dataset to make it more representative for an event detection task. If a real-world event is represented by one or more hashtags (Assumption 1), then we assume that at least one tweet with these hashtags is geotagged and that these hashtags are therefore already present in the original dataset. We thus consider every hashtag in the original dataset and use them one by one to query the Twitter REST API.

A query to the REST API returns an ordered batch of tweets $(t_i)_{i=1}^m$, where m is at most 100. By adjusting the query parameters—e.g. the maximum ID of the tweets—one can use multiple requests to gather tweets up to one week in the past. To make sure we only gather tweets from within Flanders, the tokens in the user location text field of every tweet in the current batch are compared to a list of regions, cities, towns and villages in Flanders, assembled using Wikipedia and manually adjusted for multilingual support. If the user location field is empty, the tweet is not considered further. We define a batch $(t_i)_{i=1}^m$ to be valid if and only if

$$\frac{|\{t_i : t_i \text{ in Flanders}\}|}{\text{timestamp}(t_m) - \text{timestamp}(t_1)} > \tau_1, \tag{11}$$

where τ_1 is a predefined threshold. If there are τ_2 subsequent invalid batches, all batches for the current considered hashtag are discarded. If there are τ_3 batches in total for which less than τ_4 tweets were sent in Flanders, all batches for the current considered hashtag are discarded as well. If none of these rules apply, all batches for the current hashtag are added to the dataset. When the timestamp$(\cdot)$ function is expressed in minutes, we set $\tau_1 = 1$, $\tau_2 = 12$, $\tau_3 = 25$ and $\tau_4 = 10$, as this yielded a good trade-off between execution time and quality of the data.

4.3 Collecting Events

Using the assembled data and the event definition of Section 4.1 we can assemble a ground truth for event detection in three steps. Since events are represented by one or more hashtags according to Assumption 1, we first cluster the hashtags in the tweets using a co-occurrence measure. Next we determine whether such a cluster represents an event, and finally we label the tweets corresponding with this cluster with an appropriate event label.

To assemble frequently co-occurring hashtags into clusters, a so-called co-occurrence matrix is constructed. It is a three-dimensional matrix Q that holds information on how many

times two hashtags co-occur in a tweet. Since events can only take place on one day (Assumption 2), we calculate co-occurrence on a daily basis. If hashtag k and hashtag ℓ co-occur $a_{k,\ell,d}$ times on day d, then

$$\forall k, \ell, d\colon Q_{k,\ell,d} = \frac{a_{k,\ell,d}}{\sum_i a_{k,i,d}}. \tag{12}$$

To cluster co-occurring hashtags we adopt the standard DBSCAN clustering algorithm. This is an online clustering algorithm that requires two thresholds to be set: the minimum number of hashtags $\min_h$ per cluster and a minimum similarity measure ϵ between two hashtags above which the two hashtags reside in the same ϵ-neighbourhood. The similarity measure between hashtags k and ℓ on day d is defined as

$$\mathrm{sim}_{k,\ell,d} = \frac{Q_{k,\ell,d} + Q_{\ell,k,d}}{2}. \tag{13}$$

If we run DBSCAN for every day in the dataset, we obtain a collection of clusters of sufficiently co-occurring hashtags on the same day.

A lot of these clusters however do not represent a real-world event. Hashtags such as #love or #followme do not exhibit event-specific characteristics, such as an isolated, statistically significant peak in tweet volume per minute, but can rather be seen as near-constant noise in the Twitter feed. In order to identify the hashtags that do represent events and to filter out the noise, we follow a peak detection strategy. For this purpose we treat each cluster of hashtags separately, and we refer to the hashtags in these clusters as 'event hashtags'. With each cluster C we associate all the tweets that were sent on the same day and that contain one or more of the event hashtags in this cluster. We gather them in a set T_C. After sorting the tweets in T_C according to their timestamp, we calculate how many tweets are sent in every timeslot of five minutes, which makes up for a sequence $(v_{C,i})_{i=1}^{n}$ of tweet volumes, with n the number of time slots. We define that some v_{C,i^*} is an isolated peak in the sequence $(v_{C,i})$ if and only if

$$v_{C,i^*} \geq \theta_1 \wedge \forall i \neq i^*\colon v_{C,i^*} \geq v_{C,i} + \theta_2, \tag{14}$$

with θ_1 and θ_2 predefined thresholds. Only if one such isolated peak exists (Assumption 3), we label all tweets t in T_C with the same unique event label e_t and add them to the ground truth. Since we used the event hashtags from C to construct this event, we have to remove all event hashtags in C from the tweets in T_C, otherwise the tweets themselves would already reflect the nature of the events in the ground truth.

With this procedure it is however likely that some tweets will belong to multiple events, but only get one event label. This is possible if a tweet contains multiple event hashtags that belong to different event hashtag clusters. We therefore alter the ground truth in which every tweet t corresponding to an event is associated with a set of event labels E_t instead of only one label. Of course, for the majority of these tweets, this set will only contain one event label.

In our final implementation we set $\min_h = 1$, $\epsilon = 0.3$, $\theta_1 = 10$ and $\theta_2 = 5$. These values were chosen empirically, such that, with these parameters, clusters of co-occurring hashtags are rarely bigger than three elements. After manual inspection and filtering, the final dataset contains 322 different events adding up to a total of 63,067 tweets. We assign $^2/_3$ of the events to a training set and $^1/_3$ to a test set, leading to 29,844 tweets in the training set and 33,223 in the test set.

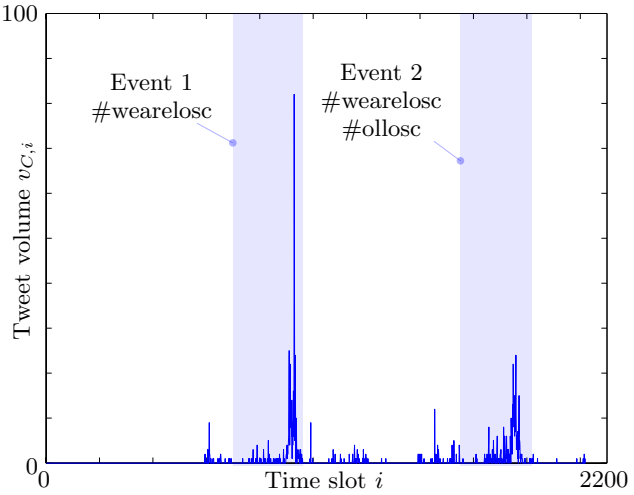


Figure 1: Plot of tweet volume in function of time slot for two example events in the dataset, with their associated hashtags.

Figure 1 shows a plot of the tweet volume in function of time slot for two events in the dataset. The plot only covers the first week in the dataset. The events are two football games of the French team LOSC Lille—which is a city very near Flanders, and therefore shows up in our dataset. The first event is characterised by the single hashtag #wearelosc, and the second event by two hashtags: #wearelosc and #ollosc. Our algorithm detects the peaks in tweet volume during the games, and since only one significant peak exists per day, we assign the same event label to all tweets with the associated hashtags sent during that day.

The final dataset is made available at the earlier mentioned URL. We provide for every tweet its tweet ID, timestamp, corresponding event labels and event hashtags, and whether it belongs to either the training or test set. Due to Twitter's restrictions, we cannot directly provide the text of all tweets.

5. RESULTS

5.1 Performance Measures

To assess the performance of the clustering algorithms, we report our results in terms of precision P, recall R and F_1 measure, as defined in [3, 15], and restated here:

$$P = \frac{1}{|T|} \sum_{t \in T} \frac{|S_t \cap \{t'\colon e_{t'} = e_t\}|}{|S_t|}, \tag{15}$$

$$R = \frac{1}{|T|} \sum_{t \in T} \frac{|S_t \cap \{t'\colon e_{t'} = e_t\}|}{|\{t'\colon e_{t'} = e_t\}|}, \tag{16}$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}, \tag{17}$$

in which T stands for the total dataset of tweets. When tweets can have multiple event labels, these definitions however do not apply any more. We therefore alter them as

	Precision	Recall	F_1-measure
Baseline	47.12%	35.35%	40.40%
Semantics-driven	52.80%	30.60%	38.74%
Hashtag semantics	48.62%	36.97%	42.00%
Baseline (multi)	64.96%	36.36%	46.62%
Semantics-driven (multi)	69.27%	31.47%	43.28%
Hashtag semantics (multi)	64.06%	37.77%	47.52%

Table 1: Using hashtag-level semantics clearly outperforms baseline and plain semantics-driven clustering.

	Purity	Number of events
Baseline	61.29%	409
Semantics-driven	64.76%	662
Hashtag semantics	61.15%	441
Baseline (multi)	75.51%	409
Semantics-driven (multi)	77.74%	662
Hashtag semantics (multi)	73.72%	441

Table 2: A comparison of baseline, plain semantics-driven clustering and hashtag semantics in terms of purity and number of event clusters.

follows:

$$P = \frac{1}{|T|} \sum_{t \in T} \max_e \frac{|S_t \cap \{t' : e \in E_{t'} \wedge e \in E_t\}|}{|S_t|}, \quad (18)$$

$$R = \frac{1}{|T|} \sum_{t \in T} \max_e \frac{|S_t \cap \{t' : e \in E_{t'} \wedge e \in E_t\}|}{|\{t' : e \in E_{t'} \wedge e \in E_t\}|}. \quad (19)$$

Note that Equations (18) and (19) imply Equations (15) and (16) if there is only one event label per tweet.

We will also use purity as an indicator of the quality of the event clusters we obtain. We have chosen the definition of purity as in [11] and adapted it to our context as follows:

$$\text{purity} = \frac{1}{|T|} \sum_{t \in T} \max_e \frac{|S_t \cap \{t' : e = e_{t'}\}|}{|S_t|}. \quad (20)$$

It is a measure that is closely related to precision. For multiple event labels, we alter this measure to the following expression:

$$\text{purity} = \frac{1}{|T|} \sum_{t \in T} \max_e \frac{|S_t \cap \{t' : e \in E_{t'}\}|}{|S_t|}. \quad (21)$$

5.2 Results

We now discuss the results of the algorithms explained in Section 3 with the use of the dataset constructed in Section 4. In the algorithms we make use of a set A of active event clusters, which become inactive after some time period. We could for example use an exponential decay function to model the time after which a cluster becomes inactive since the last tweet was added. Using Assumption 2 however we can use a much simpler method: when a new day begins, all event clusters are removed from A and thus become inactive. This way we start with an empty set A of active clusters every midnight.

For the semantics-driven clustering algorithm we assign the tweets to 10 TwitterLDA topics using the standard parameters proposed in [23] and 500 iterations of Gibbs sampling. Table 1 shows the results of the baseline algorithm, the semantics-driven algorithm and the hashtag-level semantics approach, both for one event label and multiple event labels per tweet. Note that, since we have removed the event hashtags from the tweets in the ground truth, the hashtag-level semantics approach does not use any implicit or explicit information about the nature of the events.

We note that the hashtag-level semantics approach outperforms the baseline clustering algorithm, with an increase of 1.6 percentage points in F_1-measure for single event labels. In terms of precision and recall, hashtag-level semantics performs better in both metrics than baseline in the single label case (significant improvement, $p < 0.001$ in t-test). When using multiple event labels per tweet, precision is decreased by 0.9 percentage points, but raises recall with 1.4 percentage points, leading to an increase of F_1-measure by 0.9 percentage points.

Compared to the standard semantics-driven algorithm we do 6 percentage points better in recall, but 4 percentage point worse in precision for single event labels. Hashtag-level semantic clustering seems to manage to account for the substantial loss in recall that occurs when using the basic semantics-driven method, but lacks in precision; the precision is however still 1.5 percentage points better than the baseline algorithm. The plain semantics-driven approach is 1.7 percentage points worse than baseline in terms of F_1-measure, but provides much more precision by sacrificing in recall. For multiple event labels the differences are even more pronounced between the standard semantics approach and the other algorithms. The former performs 3.3 percentage points worse in F_1-measure compared to baseline, and 4.2 percentage points worse compared to hashtag semantics. Using multiple event labels, the plain semantics-driven algorithm however has a much higher precision than baseline and hashtag semantics.

To assess the significance of the differences in F_1 measure between our three systems, we used a Bayesian technique suggested by Goutte et al. [9]. First we estimated the true positive, false positive and false negative numbers for the three systems. Next we sampled 10,000 gamma variates from the proposed distribution for F_1 for these systems and calculated the probability of one system being better than another system. We repeated this process 10,000 times. Hashtag semantics resulted in a higher F_1 measure in 99.99% of the cases; our results are thus a significant improvement over baseline. By contrast, the plain semantics-driven approach is significantly worse than baseline, also in 99.99% of the cases. Concerning multiple event labels, the hashtag semantics approach is better in 98.5% of the cases than baseline, which is also a significant improvement—although less than in the single event label case.

We also compare our three approaches in terms of cluster purity and the number of detected event clusters. These numbers are shown in Table 2. We see that the purity of the clusters in the plain semantics-driven approach is higher than baseline and hashtag semantics, but the number of detected event clusters is even substantially larger. This explains the high precision and low recall of the semantics-

driven algorithm. The purity of baseline and hashtag semantics is almost equal, but the latter approach discerns more events than baseline, thereby explaining the slight increase in precision and recall for the hashtag semantics approach compared to baseline. Concerning multiple event labels, the purity increases significantly compared to single event labels. Since the number of detected events remains the same, this explains the substantial increase in precision for the multi-label procedure.

5.3 An Illustrative Example

As a matter of example, consider the tweet *"we are ready #belgianreddevils via @sporza"*. This tweet was sent on the occasion of a football game between Belgium and Andorra—the Belgian players are called Red Devils and the airing television channel was Sporza. Since most tweets on this football game were sent in Dutch or French, the baseline clustering approach is not able to put this tweet in the correct cluster, but rather in a cluster in which most tweets are in English. This tweet is however related to a sports-specific topic, so that in both the semantics approaches the tweet is assigned to a correct cluster. It is clear that the hashtag #belgianreddevils has something to do with sports—and in particular a football game of the Belgian national team—but there exist tweets that contain this hashtag and that have not been categorized into the sports category by the TwitterLDA algorithm. For example the tweet *"met 11 man staan verdedigen, geweldig! #belgiumreddevils"* (which translates to "defending with 11 men, fantastic!") belongs to a more general category. This shows that calculating semantic topics on tweet level results in a fine-grained, but also more noisy assignment of these topics, which is reflected in the number of detected events shown in Table 2. By assigning the semantic topics on hashtag level however, all tweets with the hashtag #belgianreddevils will eventually belong to the sports category. It will result in a coarser, less detailed assignment of the topics, resulting in a more accurate event detection, and fewer detected events.

6. CONCLUSION

We developed two semantics-based extensions to the single-pass baseline clustering algorithm as used by Becker et al. to detect events in Twitter streams. In this we used semantic information about the tweets to drive the event detection. For this purpose we assigned a topic label to every tweet using the TwitterLDA algorithm. To evaluate the performance of the algorithms we semi-automatically developed a ground truth using a hashtag clustering and peak detection strategy, to aid the manual labelling of tweets with events. When using the topic labels at the level of individual tweets, the algorithm performs significantly worse than baseline. When however gathering the semantic labels of the tweets on a coarser, hashtag level we get a significant gain over baseline. We can conclude that high-level semantic information can indeed improve new and existing event detection and clustering algorithms.

7. ACKNOWLEDGMENTS

Cedric De Boom is funded by a Ph.D. grant of Ghent University, Special Research Fund (BOF).
Steven Van Canneyt is funded by a Ph.D. grant of the Agency for Innovation by Science and Technology in Flanders (IWT).

8. REFERENCES

[1] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes. Sensing Trending Topics in Twitter. *Multimedia, IEEE Transactions on*, 2013.

[2] H. Becker, M. Naaman, and L. Gravano. Event Identification in Social Media. In *WebDB 2009: Twelfth International Workshop on the Web and Databases*, 2009.

[3] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *WSDM '10: Third ACM international conference on Web search and data mining*, 2010.

[4] H. Becker, M. Naaman, and L. Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter. In *ICWSM 2011: International AAAI Conference on Weblogs and Social Media*, 2011.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Machine Learning*, 2003.

[6] L. Chen and A. Roy. Event detection from flickr data through wavelet-based spatial analysis. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, 2009.

[7] F. Chierichetti, J. Kleinberg, R. Kumar, M. Mahdian, and S. Pandey. Event Detection via Communication Pattern Analysis. In *ICWSM '14: International Conference on Weblogs and Social Media*, 2014.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning Research*, 2008.

[9] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *ECIR'05: Proceedings of the 27th European conference on Advances in Information Retrieval Research*, 2005.

[10] Y. Liu, A. Niculescu-Mizil, and W. Gryc. Topic-link LDA: joint models of topic and author community. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

[11] C. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.

[12] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. 2013.

[13] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. In *IUI '12: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, 2012.

[14] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to Twitter. In *HLT '10: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[15] T. Reuter and P. Cimiano. Event-based classification of social media streams. In *ICMR '12: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, 2012.

[16] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from twitter. In *KDD '12:*

Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th international conference on World wide web*, 2010.

[18] G. Stilo and P. Velardi. Time Makes Sense: Event Discovery in Twitter Using Temporal Similarity. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, 2014.

[19] S. Van Canneyt, S. Schockaert, and B. Dhoedt. Estimating the Semantic Type of Events Using Location Features from Flickr. In *SIGSPATIAL '14*, 2014.

[20] Y. Wang, E. Agichtein, and M. Benzi. TM-LDA: efficient online modeling of latent topic transitions in social media. In *KDD '12: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012.

[21] J. Weng, Y. Yao, E. Leonardi, and B.-S. Lee. Event Detection in Twitter. In *ICWSM '11: International Conference on Weblogs and Social Media*, 2011.

[22] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 2012.

[23] W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li. Topical keyphrase extraction from Twitter. In *HLT '11: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter

Bo Wang, Arkaitz Zubiaga, Maria Liakata, Rob Procter
Department of Computer Science
University of Warwick
Coventry, UK
{bo.wang,a.zubiaga,m.liakata,rob.procter}@warwick.ac.uk

ABSTRACT

Social spam produces a great amount of noise on social media services such as Twitter, which reduces the signal-to-noise ratio that both end users and data mining applications observe. Existing techniques on social spam detection have focused primarily on the identification of spam accounts by using extensive historical and network-based data. In this paper we focus on the detection of spam tweets, which optimises the amount of data that needs to be gathered by relying only on tweet-inherent features. This enables the application of the spam detection system to a large set of tweets in a timely fashion, potentially applicable in a real-time or near real-time setting. Using two large hand-labelled datasets of tweets containing spam, we study the suitability of five classification algorithms and four different feature sets to the social spam detection task. Our results show that, by using the limited set of features readily available in a tweet, we can achieve encouraging results which are competitive when compared against existing spammer detection systems that make use of additional, costly user features. Our study is the first that attempts at generalising conclusions on the optimal classifiers and sets of features for social spam detection over different datasets.

Categories and Subject Descriptors

I.5.4 [**Computing Methodologies**]: Pattern Recognition—*Applications*; J.4 [**Computer Application**]: Social and behavioural sciences

General Terms: Experimentation

Keywords: spam detection, classification, social media, microblogging

1. INTRODUCTION

Social networking spam, or social spam, is increasingly affecting social networking websites, such as Facebook, Pinterest and Twitter. According to a study by the social media security firm Nexgate [14], social media platforms experienced a 355% growth of social spam during the first half of 2013. Social spam can reach a surprisingly high visibility even with a simple bot [1], which detracts from a company's social media presence and damages their social marketing ROI (Return On Investment). Moreover, social spam exacerbates the amount of unwanted information that average social media users receive in their timeline, and can occasionally even affect the physical condition of vulnerable users through the so-called "*Twitter psychosis*" [7].

Social spam has different effects and therefore its definition varies across major social networking websites. One of the most popular social networking services, Twitter, has published their definition of spamming as part of their "The Twitter Rules" [1] and provided several methods for users to report spam such as tweeting "*@spam @username*" where *@username* will be reported as a spammer. While as a business, Twitter is also generous with mainline bot-level access [2] and allows some level of advertisements as long as they do not violate "The Twitter Rules". In recent years we have seen Twitter being used as a prominent knowledge base for discovering hidden insights and predicting trends from finance to public sector, both in industry and academia. The ability to sort out the signal (or the information) from Twitter noise is crucial, and one of the biggest effects of Twitter spam is that it significantly reduces the signal-to-noise ratio. Our work on social spam is motivated by the initial attempts at harvesting a Twitter corpus around a specific topic with a set of predefined keywords [21]. This led to the identification of a large amount of spam within those datasets. The fact that certain topics are trending and therefore many are tracking its contents encourages spammers to inject their spam tweets using the keywords associated with these topics to maximise the visibility of their tweets. These tweets produce a significant amount of noise both to end users who follow the topic as well as to tools that mine Twitter data.

In previous works, the automatic detection of Twitter spam has been addressed in two different ways. The first way is to tackle the task as a user classification problem, where a user can be deemed either a spammer or a non-spammer. This approach, which has been used by the majority of the works in the literature so far (see e.g., [18], [2], [11], [8], [20] and [5]), makes use of numerous features that need to gather historical details about a user, such as tweets that a user posted in the past to explore what they usually

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

[1] https://support.twitter.com/articles/18311-the-twitter-rules
[2] http://www.newyorker.com/tech/elements/the-rise-of-twitter-bots

tweet about, or how the number of followers and followings of a user has evolved in recent weeks to discover unusual behaviour. While this is ideal as the classifier can make use of extensive user data, it is often unfeasible due to restrictions of the Twitter API. The second, alternative way, which has not been as common in the literature (see e.g., [2]), is to define the task as a tweet classification problem, where a tweet can be deemed spam or non-spam. In this case, the classification task needs to assume that only the information provided within a tweet is available to determine if it has to be categorised as spam. Here, we delve into this approach to Twitter spam classification, studying the categorisation of a tweet as spam or not from its inherent features. While this is more realistic for our scenario, it presents the extra challenge that the available features are rather limited, which we study here.

In this work, after discussing the definition of social spam and reviewing previous research in Twitter spam detection, we present a comparative study of Twitter spam detection systems. We investigate the use of different features inherent to a tweet so as to identify the sets of features that do best in categorising tweets as spam or not. Our study compares five different classification algorithms over two different datasets. The fact that we test our classifiers on two different datasets, collected in different ways, enables us to validate the results and claim repeatability. Our results suggest a competitive performance can be obtained using tree-based classifiers for spam detection even with only tweet-inherent features, as comparing to the existing spammer detection studies. Also the combination of different features generally lead to an improved performance, with User feature + Bi & Tri-gram (Tf) having the best results for both datasets.

2. SOCIAL SPAM

The detection of spam has now been studied for more than a decade since email spam [4]. In the context of email messages, spam has been widely defined as "unsolicited bulk email" [3]. The term "spam" has then been extended to other contexts, including "social spam" in the context of social media. Similarly, social spam can be defined as the "unwanted content that appears in online social networks". It is, after all, the noise produced by users who express a different behavior from what the system is intended for, and has the goal of grabbing attention by exploiting the social networks' characteristics, including for instance the injection of unrelated tweet content in timely topics, sharing malicious links or fraudulent information. Social spam hence can appear in many different forms, which poses another challenge of having to identify very different types of noise for social spam detection systems.

2.1 Social Spammer Detection

As we said before, most of the previous work in the area has focused on the detection of users that produce spam content (i.e., spammers), using historical or network features of the user rather than information inherent to the tweet. Early work by [18], [2] and [11] put together a set of different features that can be obtained by looking at a user's previous behaviour. These include some aggregated statistics from a user's past tweets such as average number of hashtags, average number of URL links and average number of user mentions that appear in their tweets. They combine these with other non-historical features, such as number of followers, number of followings and age of the account, which can be obtained from a user's basic metadata, also inherent to each tweet they post. Some of these features, such as the number of followers, can be gamed by purchasing additional followers to make the user look like a regular user account.

Lee et al. [8] and Yang et al. [20] employed different techniques for collecting data that includes spam (more details will be discussed in Section 3.1) and performed comprehensive studies of the spammers' behaviour. They both relied on the tweets posted in the past by the users and their social networks, such as tweeting rate, following rate, percentage of bidirectional friends and local clustering coefficient of its network graph, aiming to combat spammers' evasion tactics as these features are difficult or costly to simulate. Ferrara et al. [5] used network, user, friends, timing, content and sentiment features for detecting Twitter bots, their performance evaluation is based on the social honeypots dataset (from [8]). Miller et al. [12] treats spammer detection as an anomaly detection problem as clustering algorithms are proposed and such clustering model is built on normal Twitter users with outliers being treated as spammers. They also propose using 95 uni-gram counts along with user profile attributes as features. The sets of features utilised in the above works require the collection of historical and network data for each user, which do not meet the requirements of our scenario for spam detection.

2.2 Social Spam Detection

Few studies have addressed the problem of spam detection. Santos et al. [16] investigated two different approaches, namely compression-based text classification algorithms (i.e. Dynamic Markov compression and Prediction by partial matching) and using "bag of words" language model (also known as uni-gram language model) for detecting spam tweets. Martinez-Romo and Araujo [10] applied Kullback-Leibler Divergence and examined the difference of language used in a set of tweets related to a trending topic, suspicious tweets (i.e. tweets that link to a web page) and the page linked by the suspicious tweets. These language divergence measures were used as their features for the classification. They used several URL blacklists for identifying spam tweets from their crawled dataset, therefore each one of their labelled spam tweets contains a URL link, and is not able to identify other types of spam tweets. In our studies we have investigated and evaluated the discriminative power of four feature sets on two Twitter datasets (which were previously in [8] and [20]) using five different classifiers. We examine the suitability of each of the features for the spam classification purposes. Comparing to [10] our system is able to detect most known types of spam tweet irrespective of having a link or not. Also our system does not have to analyze a set of tweets relating to each topic (which [10] did to create part of their proposed features) or external web page linked by each suspicious tweet, therefore its computation cost does not increase dramatically when applied for mass spam detection with potentially many different topics in the data stream.

The few works that have dealt with spam detection are mostly limited in terms of the sets of features that they studied, and the experiments have been only conducted in a single dataset (except in the case of [10], where very limited evaluation was conducted on a new and smaller set of tweets), which does not allow for generalisability of the re-

sults. To the best of our knowledge, our work is the first study that evaluates a wide range of tweet-inherent features (namely user, content, n-gram and sentiment features) over two different datasets, obtained from [8] and [20] and with more than 10,000 tweets each, for the task of spam detection. The two datasets were collected using completely different approaches (namely deploying social honeypots for attracting spammers; and checking malicious URL links), which helps us learn more about the nature of social spam and further validate the results of different spam detection systems.

3. METHODOLOGY

In this section we describe the Twitter spam datasets we used, the text preprocessing techniques that we performed on the tweets, and the four different feature sets we used for training our spam vs non-spam classifier.

3.1 Datasets

A labelled collection of tweets is crucial in a machine learning task such as spam detection. We found no spam dataset which is publicly available and specifically fulfils the requirements of our task. Instead, the datasets we obtained include Twitter users labelled as spammers or not. For our work, we used the latter, which we adapted to our purposes by taking out the features that would not be available in our scenario of spam detection from tweet-inherent features. We used two spammer datasets in this work, which have been created using different data collection techniques and therefore is suitable to our purposes of testing the spam classifier in different settings. To accomodate the datasets to our needs, we sample one tweet for each user in the dataset, so that we can only access one tweet per user and cannot aggregate several tweets from the same user or use social network features. In what follows we describe the two datasets we use.

Social Honeypot Dataset: Lee et al. [8] created and manipulated (by posting random messages and engaging in none of the activities of legitimate users) 60 social honeypot accounts on Twitter to attract spammers. Their dataset consists of 22,223 spammers and 19,276 legitimate users along with their most recent tweets. They used Expectation-Maximization (EM) clustering algorithm and then manually grouped their harvested users into 4 categories: duplicate spammers, duplicate @ spammers, malicious promoters and friend infiltrators. **1KS-10KN Dataset:** Yang et al. [20] defines a tweet that contains at least one malicious or phishing URL as a spam tweet, and a user whose spam ratio is higher than 10% as a spammer. Therefore their dataset which contains 1,000 spammers and 10,000 legitimate users, represents only one major type of spammers (as discussed in their paper).

We used *spammer vs. legitimate user* datasets from [8] and [20]. After removing duplicated users and the ones that do not have any tweets in the dataset we randomly selected one tweet from each spammer or legitimate user to create our labelled collection of *spam vs. legitimate tweets*, in order to avoid overfitting and reduce our sampling bias. The resulting datasets contain 20,707 spam tweets and 19,249 normal tweets (named Social Honeypot dataset, as from [8]), and 1,000 spam tweets and 9,828 normal tweets (named 1KS-10KN dataset, as from [20]) respectively.

3.2 Data Preprocessing

Before we extract the features to be used by the classifier from each tweet, we apply a set of preprocessing techniques to the content of the tweets to normalise it and reduce the noise in the classification phase. The preprocessing techniques include decoding HTML entities, and expanding contractions with apostrophes to standard spellings (e.g. "I'm" -> "I am"). More advanced preprocessing techniques such as spell-checking and stemming were tested but later discarded given the minimal effect we observed in the performance of the classifiers.

For the specific case of the extraction of sentiment-based features, we also remove hashtags, links, and user mentions from tweet contents.

3.3 Features

As spammers and legitimate users have different goals in posting tweets or interacting with other users on Twitter, we can expect that the characteristics of spam tweets are quite different to the normal tweets. The features inherent to a tweet include, besides the tweet content itself, a set of metadata including information about the user who posted the tweet, which is also readily available in the stream of tweets we have access to in our scenario. We analyse a wide range of features that reflect user behaviour, which can be computed straightforwardly and do not require high computational cost, and also describe the linguistic properties that are shown in the tweet content. We considered four feature sets: (i) user features, (ii) content features, (iii) n-grams, and (iv) sentiment features.

User features include a list of 11 attributes about the author of the tweet (as seen in Table 1) that is generated from each tweet's metadata, such as reputation of the user [18], which is defined as the ratio between the number of followers and the total number of followers and followings and it had been used to measure user influence. Other candidate features, such as the number of retweets and favourites garnered by a tweet, were not used given that it is not readily available at the time of posting the tweet, where a tweet has no retweets or favourites yet.

Content features capture the linguistic properties from the text of each tweet (Table 1) including a list of content attributes and part-of-speech tags. Among the 17 content attributes, number of spam words and number of spam words per word are generated by matching a popular list of spam words [3]. Part-of-speech (or POS) tagging provides syntactic (or grammatical) information of a sentence and has been used in the natural language processing community for measuring text informativeness (e.g. Tan et al. [17] used POS counts as a informativeness measure for tweets). We have used a Twitter-specific tagger [6], and in the end our POS feature consists of uni-gram and 2-skip-bi-gram representations of POS tagging for each tweet in order to capture the structure and therefore informativeness of the text. We also used Stanford tagger with standard Penn Tree tags, which makes very little difference in the classification results.

N-gram models have long been used in natural language processing for various tasks including text classification. Although it is often criticized for its lack of any explicit representation of long range or semantic dependency, it is surpris-

[3]https://github.com/splorp/wordpress-comment-blacklist/blob/master/blacklist.txt

User features	Content features
Length of profile name	Number of words
Length of profile description	Number of characters
Number of followings (FI)	Number of white spaces
Number of followers (FE)	Number of capitalization words
Number of tweets posted	Number of capitalization words per word
Age of the user account, in hours (AU)	Maximum word length
Ratio of number of followings and followers (FE/FI)	Mean word length
Reputation of the user (FE/(FI + FE))	Number of exclamation marks
Following rate (FI/AU)	Number of question marks
Number of tweets posted per day	Number of URL links
Number of tweets posted per week	Number of URL links per word
N-grams	Number of hashtags
Uni + bi-gram or bi + tri-gram	Number of hashtags per word
	Number of mentions
Sentiment features	Number of mentions per word
Automatically created sentiment lexicons	Number of spam words
Manually created sentiment lexicons	Number of spam words per word
	Part of speech tags of every tweet

Table 1: List of features

ingly powerful for simple text classification with reasonable amount of training data. In order to give the best classification result while being computationally efficient we have tried uni + bi-gram or bi + tri-gram with binary (i.e. 1 for feature presence while 0 for absence), term-frequency (tf) and tf-idf (i.e. Term Frequency times Inverse Document Frequency) techniques.

Sentiment features: Ferrara et al. [5] used tweet-level sentiment as part of their feature set for the purpose of detecting Twitter bots. We have used the same list of lexicons from [13] (which has been proved of achieving top performance in the Semeval-2014 Task 9 Twitter sentiment analysis competition) for generating our sentiment features, including manually generated sentiment lexicons: AFINN lexicon [15] , Bing Liu lexicon [9], MPQA lexicon [19]; and automatically generated sentiment lexicons: NRC Hashtag Sentiment lexicon [13] and Sentiment140 lexicon [13].

4. EVALUATION

4.1 Selection of Classifier

During the classification and evaluation stage, we tested 5 classification algorithms implemented using scikit-learn[4]: Bernoulli Naive Bayes, K-Nearest Neighbour (KNN), Support Vector Machines (SVM), Decision Tree, and Random Forests. These algorithms were chosen as being the most commonly used in the previous research on spammer detection. We evaluate using the standard information retrieval metrics of recall (R), precision (P) and F1-measure. Recall in this case refers to the ratio obtained from diving the number of correctly classified spam tweets (i.e. True Positives) by the number of tweets that are actually spam (i.e. True Positives + False Negatives). Precision is the ratio of the number of correctly classified spam tweets (i.e. True Positives) to the total number of tweets that are classified as spam (i.e. True Positives + False Positives). F1-measure can be interpreted as a harmonic mean of the precision and recall, where its score reaches its best value at 1 and worst at 0. It is defined as:

$$F1 = 2 * (precision * recall)/(precision + recall)$$

In order to select the best classifier for our task, we have used a subset of each dataset (20% for 1KS-10KN dataset and 40% for Social Honeypot dataset, due to the different sizes of the two datasets) to run a 10-fold cross validation for optimising the hyperparameters of each classifier. By doing so it minimises the risk of over-fitting in model selection and hence subsequent selection bias in performance evaluation. Such optimisation was conducted using all 4 feature sets (each feature was normalised to fit the range of values [-1, 1]; we also selected 30% of the highest scoring features using Chi Square for tuning SVM as computationally it is more efficient and gives better classification results). Then we evaluated our algorithm on the rest of the data (i.e. 80% for 1KS-10KN dataset and 60% for Social Honeypot dataset), again using all 4 feature sets in a 10-fold cross validation setting (same as in grid-search, each feature was normalised and Chi square feature selection was used for SVM).

As shown in Table 2, tree-based classifiers achieved very promising performances, among which Random Forests out-

[4]http://scikit-learn.org/

perform all the others when we look at the F1-measure. This outperformance occurs especially due to the high precision values of 99.3% and 94.1% obtained by the Random Forest classifier. While Random Forests show a clear superiority in terms of precision, its performance in terms of recall varies for the two datasets; it achieves high recall for the Social Honeypot dataset, while it drops substantially for the 1KS-10KN dataset due to its approximate 1:10 spam/non-spam ratio. These results are consistent with the conclusion of most spammer detection studies; our results extend this conclusion to the spam detection task.

When we compare the performance values for the different datasets, it is worth noting that with the Social Honeypot dataset the best result is more than 10% higher than the best result in 1KS-10KN dataset. This is caused by the different spam/non-spam ratios in the two datasets, as the Social Honeypot dataset has a roughly 50:50 ratio while in 1KS-10KN it is roughly 1:10 which is a more realistic ratio to reflect the amount of spam tweets existing on Twitter (In Twitter's 2014 Q2 earnings report it says that less than 5% of its accounts are spam[5], but independent researchers believe the number is higher). In comparison to the original papers, [8] reported a best 0.983 F1-score and [20] reported a best 0.884 F1-score. Our results are only about 4% lower than their results, which make use of historical and network-based data, not readily available in our scenario. Our results suggest that a competitive performance can also be obtained for spam detection where only tweet-inherent features can be used.

4.2 Evaluation of Features

We trained our best classifier (i.e. Random Forests) with different feature sets, as well as combinations of the feature sets using the two datasets (i.e. the whole corpora), and under a 10-fold cross validation setting. We report our results in Table 3. As seen in 1KS-10KN dataset, the F1-measure for different feature sets ranges from 0.718 to 0.820 when using a single feature set. All feature set combinations except C + S (content + sentiment feature) perform higher than 0.810 in terms of F1-measure, reflecting that feature combinations have more discriminative power than a single feature set.

For the Social Honeypot dataset, we can clearly see User features (U) having the most discriminative power as it has a 0.940 F1-measure. Results without using User features (U) have significantly worse performance, and feature combinations with U give very little improvement with respect to the original 0.940 (except for U + Uni & Bi-gram (Tf) + S). This means U is dominating the discriminative power of these feature combinations and other feature sets contribute very little in comparison to U. This is potentially caused by the data collection approach (i.e. by using social honeypots) adopted by [8], which resulted in the fact that most spammers that they attracted have distinguishing user profile information compared to the legitimate users. On the other hand, Yang et al. [20] checked malicious or phishing URL links for collecting their spammer data, and this way of data collection gives more discriminative power to Content and N-gram features than [8] does (although U is still a very significant feature set in 1KS-10KN). Note that U + Bi & Tri-gram (Tf) resulted in the best performance in both datasets, showing that these two feature sets are the most beneficial to each other irrespective of the different nature of datasets.

Another important aspect to take into account when choosing the features to be used is the computation time, especially when one wants to apply the spam classifier in real-time. Table 4 shows a efficiency comparison for generating each feature from 1000 tweets, using a machine with 2.8 GHz Intel Core i7 processor and 16 GB memory. Some of the features, such as the User features, can be computed quickly and require minimal computational cost, as most of these features can be straightforwardly inferred from a tweet's metadata. Other features, such as N-grams and part-of-speech counts (from Content features), can be affected by the size of the vocabulary in the training set. On the other hand, some of the features are computationally more expensive, and therefore worth studying their applicability. This is the case of Sentiment features, which require string matching between our training documents and a list of lexica we used. We keep the sentiment features since they have shown added value in the performance evaluation of feature set combinations. Similarly, Content features such as *Number of spam words* and *Number of spam words per word* also require string matching between our training documents and a dictionary containing 11,529 spam words. However, given that the latter did not provide significant improvements in terms of accuracy, most probably because the spam words were extracted from blogs, we conclude that *Number of spam words* and *Number of spam words per word* can be taken out from the representation for the sake of the classifier's efficiency.

5. DISCUSSION

Our study looks at different classifiers and feature sets over two spam datasets to pick the settings that perform best. First, our study on spam classification buttresses previous findings for the task of spammer classification, where Random Forests were found to be the most accurate classifier. Second, our comparison of four feature sets reveals the features that, being readily available in each tweet, perform best in identifying spam tweets. While different features perform better for each of the datasets when using them alone, our comparison shows that the combination of different features leads to an improved performance in both datasets. We believe that the use of multiple feature sets increases the possibility to capture different spam types, and makes it more difficult for spammers to evade all feature sets used by the spam detection system. For example spammers might buy more followers to look more legitimate but it is still very likely that their spam tweet will be detected as its tweet content will give away its spam nature.

Due to practical limitations, we have generated our spam vs. non-spam data from two spammer vs. non-spammer datasets that were collected in 2011. For future work, we plan to generate a labelled spam/non-spam dataset which was crawled in 2014. This will not only give us a purpose-built corpus of spam tweets to reduce the possible effect of sampling bias of the two datasets that we used, but will also give us insights on how the nature of Twitter spam changes over time and how spammers have evolved since 2011 (as spammers do evolve and their spam content are manipulated to look more and more like normal tweet). Furthermore we will investigate the feasibility of cross-dataset spam classification using domain adaptation methods, and also whether

[5]http://www.webcitation.org/6VyBTJ7vt

Classifier	1KS-10KN Dataset			Social Honeypot Dataset		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Bernoulli NB	0.899	0.688	0.778	0.772	0.806	0.789
KNN	0.924	0.706	0.798	0.802	0.778	0.790
SVM	0.872	0.708	0.780	0.844	0.817	0.830
Decision Tree	0.788	0.782	0.784	0.914	0.916	0.915
Random Forest	0.993	0.716	**0.831**	0.941	0.950	**0.946**

Table 2: Comparison of performance of classifiers

Feature Set	1KS-10KN Dataset			Social Honeypot Dataset		
	Precision	Recall	F-measure	Precision	Recall	F-measure
User features (U)	0.895	0.709	0.791	0.938	0.940	0.940
Content features (C)	0.951	0.657	0.776	0.771	0.753	0.762
Uni + Bi-gram (Binary)	0.930	0.725	0.815	0.759	0.727	0.743
Uni + Bi-gram (Tf)	0.959	0.715	0.819	0.783	0.767	0.775
Uni + Bi-gram (Tfidf)	0.943	0.726	0.820	0.784	0.765	0.775
Bi + Tri-gram (Tfidf)	0.931	0.684	0.788	0.797	0.656	0.720
Sentiment features (S)	0.966	0.574	0.718	0.679	0.727	0.702
U + C	0.974	0.708	0.819	0.938	0.949	**0.943**
U + Bi & Tri-gram (Tf)	0.972	0.745	**0.843**	0.937	0.949	**0.943**
U + S	0.948	0.732	0.825	0.940	0.944	0.942
Uni & Bi-gram (Tf) + S	0.964	0.721	0.824	0.797	0.744	0.770
C + S	0.970	0.649	0.777	0.778	0.762	0.770
C + Uni & Bi-gram (Tf)	0.968	0.717	0.823	0.783	0.757	0.770
U + C + Uni & Bi-gram (Tf)	0.985	0.727	0.835	0.934	0.949	0.941
U + C + S	0.982	0.704	0.819	0.937	0.948	0.942
U + Uni & Bi-gram (Tf) + S	0.994	0.720	0.834	0.928	0.946	0.937
C + Uni & Bi-gram (Tf) + S	0.966	0.720	0.824	0.806	0.758	0.782
U + C + Uni & Bi-gram (Tf) + S	0.988	0.725	0.835	0.936	0.947	0.942

Table 3: Performance evaluation of various feature set combinations

Feature set	Computation time (in seconds) for 1000 tweets
User features	0.0057
N-gram	0.3965
Sentiment features	20.9838
Number of spam words (NSW)	19.0111
Part-of-speech counts (POS)	0.6139
Content features including NSW and POS	20.2367
Content features without NSW	1.0448
Content features without POS	19.6165

Table 4: Feature engineering computation time for 1000 tweets

unsupervised approaches work well enough in the domain of Twitter spam detection.

A caveat of the approach we relied on for the dataset generation is the fact that we have considered spam tweets posted by users who were deemed spammers. This was done based on the assumption that the majority of social spam tweets on Twitter are shared by spam accounts. However, the dataset could also be complemented with spam tweets which are occasionally posted by legitimate users, which our work did not deal with. An interesting study to complement our work would be to look at these spam tweets posted by legitimate users, both to quantify this type of tweets, as well as to analyse whether they present different features from those in our datasets, especially when it comes to the user-based features as users might have different characteristics. For future work, we plan to conduct further evaluation on how our features would function for spam tweets shared by legitimate users, in order to fully understand the effects of bias of pursuing our approach of corpus construction.

6. CONCLUSION

In this paper we focus on the detection of spam tweets, solely making use of the features inherent to each tweet. This differs from most previous research works that classified Twitter users as spammers instead, and represents a real scenario where either a user is tracking an event on Twitter, or a tool is collecting tweets associated with an event. In these situations, the spam removal process cannot afford to retrieve historical and network-based features for all the tweets involved with the event, due to high number of requests to the Twitter API that this represents. We have tested five different classifiers, and four different feature sets

on two Twitter spam datasets with different characteristics, which allows us to validate our results and claim repeatability. While the task is more difficult and has access to fewer data than a spammer classification task, our results show competitive performances. Moreover, our system can be applied for detecting spam tweets in real time and does not require any feature not readily available in a tweet.

Here we have conducted the experiments on two different datasets which were originally collected in 2011. While this allows us to validate the results with two datasets collected in very different methods, our plan for future work includes the application of the spam detection system to more recent events, to assess the validity of the classifier with recent data as Twitter and spammers may have evolved.

7. REFERENCES

[1] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo. People are strange when you're a stranger: Impact and influence of bots on social networks. *CoRR*, abs/1407.8134, 2014.

[2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Proceedings of CEAS*, 2010.

[3] E. Blanzieri and A. Bryl. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29(1):63–92, 2008.

[4] X. Carreras, L. S. Marquez, and J. G. Salgado. Boosting trees for anti-spam email filtering. In *Proceedings of RANLP*. Citeseer, 2001.

[5] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *CoRR*, abs/1407.5225, 2014.

[6] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of ACL*, HLT '11, pages 42–47, Stroudsburg, PA, USA, 2011.

[7] J. Kalbitzer, T. Mell, F. Bermpohl, M. A. Rapp, and A. Heinz. Twitter psychosis: a rare variation or a distinct syndrome. *Journal of Nervous and Mental Disease*, 202(8):623, August 2014.

[8] K. Lee, B. D. Eoff, and J. Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In L. A. Adamic, R. A. Baeza-Yates, and S. Counts, editors, *ICWSM*, 2011.

[9] B. Liu. Sentiment analysis: a multifaceted problem. *IEEE Intelligent Systems*, 25(3):76–80, 2010.

[10] J. Martinez-Romo and L. Araujo. Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 40(8):2992 – 3000, 2013.

[11] M. McCord and M. Chuah. Spam detection on twitter using traditional classifiers. In J. M. A. Calero, L. T. Yang, F. G. Mármol, L. J. Garcá-Villalba, X. A. Li, and Y. W. 0002, editors, *ATC*, volume 6906 of *Lecture Notes in Computer Science*, pages 175–186. Springer, 2011.

[12] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang. Twitter spammer detection using data stream clustering. *Information Sciences*, 260(0):64 – 73, 2014.

[13] S. Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June 2013.

[14] H. Nguyen. Research report: 2013 state of social media spam. `http://nexgate.com/wp-content/uploads/2013/09/Nexgate-2013-State-of-Social-Media-Spam-Research-Report.pdf`, 2013.

[15] F. Å. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[16] I. Santos, I. Miñambres-Marcos, C. Laorden, P. Galán-García, A. Santamaría-Ibirika, and P. G. Bringas. Twitter content-based spam filtering. In *International Joint Conference SOCO'13-CISIS'13-ICEUTE'13*, pages 449–458. Springer, 2014.

[17] C. Tan, L. Lee, and B. Pang. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. *CoRR*, abs/1405.1438, 2014.

[18] A. H. Wang. Don't follow me - spam detection in twitter. In S. K. Katsikas and P. Samarati, editors, *SECRYPT*, pages 142–151. SciTePress, 2010.

[19] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[20] C. Yang, R. C. Harkreader, and G. Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *Proceedings of RAID*, RAID'11, pages 318–337, Berlin, Heidelberg, 2011. Springer-Verlag.

[21] A. Zubiaga, M. Liakata, R. Procter, K. Bontcheva, and P. Tolmie. Towards detecting rumours in social media. In *AAAI Workshop on AI for Cities*, 2015.

User Interest Modeling in Twitter with Named Entity Recognition

Deniz Karatay
METU Computer Engineering Dept.
06800 Ankara, Turkey
deniz.karatay@ceng.metu.edu.tr

Pinar Karagoz
METU Computer Engineering Dept.
06800 Ankara, Turkey
karagoz@ceng.metu.edu.tr

ABSTRACT
Considering wide use of Twitter as the source of information, reaching an interesting tweet for a user among a bunch of tweets is challenging. In this work we propose a Named Entity Recognition (NER) based user profile modeling for Twitter users and employ this model to generate personalized tweet recommendations. Effectiveness of the proposed method is shown through a set of experiments.

Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: Miscellaneous

General Terms
Theory

Keywords
Named Entity Recognition, Tweet Segmentation, Tweet Classification, Tweet Ranking, Tweet Recommendation

1. INTRODUCTION
As a service that embodies both social networking and microblogging, Twitter has become one of the most important communication channels with its ability of providing the most up-to-date and newsworthy information [6]. In this study, we present a technique for constructing user interest model, in which user interests are defined by means of relationship between the user and his friends as well as named entities extracted from tweets. We demonstrate the use of this model for tweet recommendation.

To extract information from this large volume of tweets generated by Twitter's millions of users, Named Entity Recognition (NER), which is the focus of this work, is already being used by researchers. NER can be basically defined as identifying and categorizing certain type of data (i.e. person, location, organization names, date-time and numeric expressions) in a certain type of text. On the other hand, tweets are characteristically short and noisy. Considering the fact that tweets generally include grammar mistakes, misspellings, and informal capitalization, performance of the traditional methods is incompetent on tweets and new approaches have to be generated to deal with this type of data. Recently, tweet representation based on segments in order to extract named entities has proven its validity in NER field [4, 3].

In this work, it is aimed to reduce the Twitter user's effort to access to the tweet carrying the information of interest. To this aim, a tweet recommendation method under a user interest model generated via named entities is presented. To achieve our goal, a graph based user interest model is generated via named entities extracted from user's followees' and user's own posts. In the user interest model, each included followee is ranked based on their interactions with the user via retweets and mentions, and named entities are scored via ranking of the user posting them.

2. PROPOSED METHOD
The general overview of the system architecture can also be seen in Figure 1. The method used in this study segments the tweets and generates named entity candidates. These candidates have to be validated so that they can be used as an indicator of the user's interest. In this step, Wikipedia is chosen as a reference for a segment to be a named entity, or not. Since our Tweet collection is in Turkish, Turkish Wikipedia dump published by Wikipedia is obtained.

For named entities to be extracted successfully, the informal writing style in tweets has to be handled. Generally named entities are assumed as words written in uppercase or mixed case phrases where uppercased letters are at the beginning and ending, and almost all of the studies bases on this assumption. However, capitalization is not a strong indicator in tweet-like informal texts, sometimes even misleading. To extract named entities in tweets, the effect of the informality of the tweets has to be minimized as possible. The preprocessing tasks applied can be divided into two logical group:. Pre-segmenting, and Correcting. Removal of links, hashtags, mentions, conjunctives, stop words, vocatives, slang words and elimination of punctuation are considered as presegmentation. It is assumed that parts in the texts before and after a redundant word, or a punctuation mark cannot form a named entity together, therefore every removal of a word is considered as it segments the tweet as well as punctuation does it naturally. Removal of repeating characters that are used to express a feeling such as exaggerating,

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

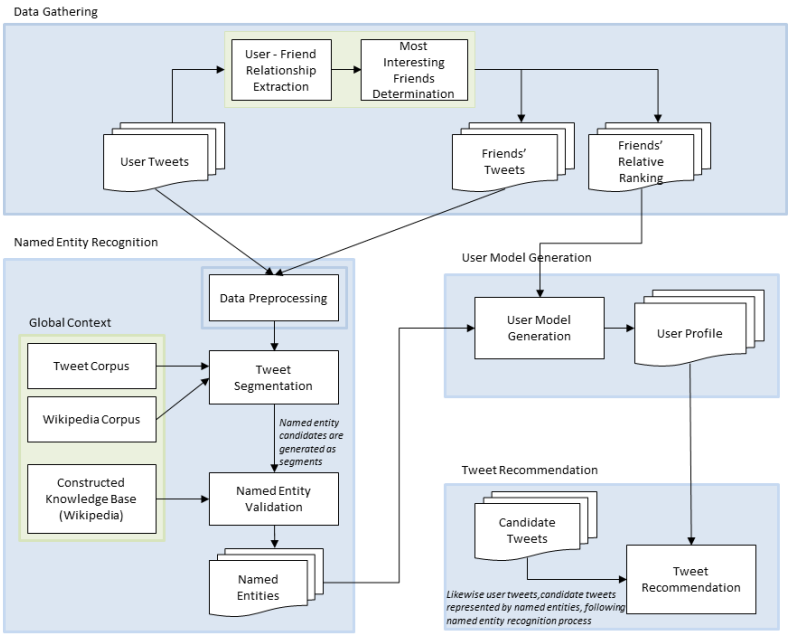


Figure 1: System Architecture

or yelling, handling mistyping and asciification related problems are considered as correcting and can be thought of conversion of tweets from informal to formal. In the following subsections, we describe the NER and user profile modeling and recommendation steps in more detail.

2.1 Finding Named Entities

In this study, the idea of segmenting a tweet text into a set of phrases, each of which appears more than random occurence [1, 4] is adopted. Therefore, a corpus serving this purpose in Turkish is needed. To this aim, *TS Corpus*, which indexes Wikipedia articles and also Tweets [5], is used. In the proposed solution, *TS Corpus* is used for gathering statistical information for various segmentation combinations by means of a dynamic programming algorithm. While collecting statistical information for segment combinations, tweet collection of *TS Corpus* is also used while computing probability of a segment to be a valid named entity, which is different from the previous studies. The knowledge base that is constructed using Turkish *Wikipedia* dump is used to validate the candidate named entities.

Segmentation constitutes the core part of named entity recognition method. The aim here is to split a tweet into consecutive segments. Each segment contains at least one word. For the optimal segmentation, the following objective function is used, where F is the *stickiness* function, t is an individual tweet, and s represents a segment.

$$\arg\max_{s_1 \ldots s_n} F(t) = \sum_{i=1}^{n} F(s_i) \tag{1}$$

Although the term *stickiness* is generally used for expressing tendency of a user to stay longer on a web page by a user, Li et. al defined it as the metric of a word group to be seen together in documents frequently, or not [4] and it is used in the same way in this study. The *stickiness* function basically measures the *stickiness* of a segment or a tweet represented based on word collocations. A low *stickiness* value of a segment means that words are not used commonly together and can be further split to obtain a more suitable word collocation. On the other hand, a high *stickiness* value of a segment indicates that words in the segment are used together often and represent a word collocation, therefore cannot be further split. In order to determine the correct segmentation, the objective function above is used, where a tweet representation with the maximum *stickiness* is chosen to be the correct segmentation. Instead of generating all possible segmentations and compute their stickiness, dynamic programming algorithm described in [4] is adapted to this study to compute stickiness values efficiently. The algorithm basically segments the longer segment, which can be tweet itself, into two segments and evaluates the *stickiness* of the resultant segments recursively. More formally, given any segment $s = w_1 w_2 ... w_n$, adjacent binary segmentations $s_1 = w_1 ... w_j$ and $s_2 = w_j + 1 ... w_n$ is obtained by satisfying the following equation.

$$\arg\max_{s_1, s_2} F(s) = F(s_1) + F(s_2) \tag{2}$$

Thus far, tweets are segmented making use of the stickiness function. In the result of this phase, tweet segments, which are candidate named entities, are obtained. These candidate named entities have to be validated whether they are real named entities or not, so that they can be used as an indicator of the user's interest. For this purpose, as explained before, Wikipedia is chosen as a reference for a segment to be a named entity, and a graph-based knowledge-base based on Wikipedia is constructed. If the segment, which is actually a candidate named entity, matches exactly with a Wikipedia title in the constructed knowledge base, then it is accepted to be a named entity. In case of inexact match, we use the

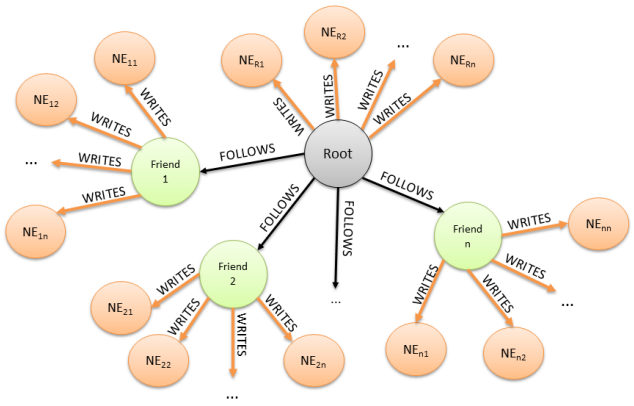


Figure 2: Structure of the User Interest Model Graph

Levenshtein distance [2] to measure the similarity of a segment to a Wikipedia title.

2.2 Generating User Interest Model based on Named Entities

At this step, named entities with their frequency counts in a tweet obtained from followees' posts, and followees' relative ranking obtained in data gathering phase is processed as shown in Figure 1. Using these data, a user interest model is generated. It is basically a graph based relationship model. Let $G = (V, E)$ be a weighted labelled graph with the node set V and edge set E. Node set V is labelled with the label set L_1 where $L_1 \in \{Root, Followee, NamedEntity\}$ and Edge set E is labelled with the label set L_2 where $L_2 \in \{Follows, Writes\}$. In other words, a user interest model graph has three types of nodes; *Root, Friend, Named Entity*, along with two types of weighted edges; *Writes, and Follows*. Weight of *Writes* edge represents the appearance count of a named entity for a followere's posts where weight of the *Follows* edge represents relative ranking of a followed. Therefore, a twitter profile is represented as *Root* node *Follows* one or many *Followees*, and a *Followee* node *Writes* one or many *Named Entities*. The structure of the graph is shown in Figure 2.

2.3 Tweet Recommendation

Determining whether a tweet is interesting or not is achieved by comparing NE representation of the tweet with the generated user interest model. This comparison results in a ranking of candidate tweets. As the first step, candidate tweets are processed to obtain their NE representations. NE representation of a tweet simply includes the NEs, and their frequency counts. In order to compare with the candidate tweet, user interest model has to be interpreted by including the ranking score factor of the friends. Every followee's named entities and their appearance counts are first multiplied with the friend's ranking, and then summed. Therefore, a set of named entities with their scores based on the user interest model is obtained. The mathematical interpretation to calculate the score of a single named entity is given in Equation 3, where SC_{NE} represents the overall score of a named entity, C represents the frequency count of a named entity for a user, n represents the count of friends included in the user interest model, RR represents the relative ranking score of a followed, and U represents the user himself. With the same approach, the final score of all of the named entities appearing in the user interest model is calculated.

$$SC_{NE} = \sum_{i=1}^{n} RR_i \cdot C_i + RR_U \cdot C_U \quad (3)$$

After overall score is calculated for all of the named entities in the user interest model, final scores for candidate tweets are calculated in the following approach: Overall score of named entities in NE representation of a candidate tweet are multiplied with the frequency count in the NE representation of itself. This operation is done for every named entity in the tweet representation, and then by summing these values, final score of a candidate tweet is obtained. If a named entity in a candidate tweet's NE representation, does not appear in the user interest model, its overall score is accepted as 0 and not taken into consideration assuming the user is not interested in the subject that particular named entity represents. Once final scores for all candidate tweets are calculated, candidate tweets are sorted in descending order, and hence, they are ranked.

$$SC_T = \sum_{i=1}^{m} SC_{NE_i} \cdot C_{NE_i} \quad (4)$$

3. EXPERIMENTAL RESULTS

To evaluate the system from recommendation point of view, two types of datasets as candidate tweets for recommendation and two types of user groups to recommend tweets are formed. The first dataset of candidate tweets, $GNRL$, is a general dataset containing 100 tweets crawled from newspapers' Twitter accounts. The second dataset, $PSNL$ is a personal dataset containing 100 tweets that are crawled from the followees of followees of the selected users. There are 10 users volunteered for this experiment where half of them are active Twitter users, whereas the other half are inactive Twitter users. *Active Users* are the users that use Twitter frequently, have retweeting and mentioning habits, and update followed list when necessary where *Inactive Users* do not post, retweet, or mention often, and do not update followee list frequently. Volunteered users are categorized on the basis of the information they provided about their Twitter usage habits.

For each user, user interest model is constructed under SCP measure on Wikipedia Corpus along with length normalization for stickiness function, which gives the best results according to the validation experiments. In addition, the best N_T and N_F values are experimentally obtained, therefore 20 followees and 10 tweets of each followed are included in the model. Candidate tweets are scored by comparing with user's model as explained in Section 2.3 and then ranked. Meanwhile, each user is asked to classify and score tweets in $GNRL$ and $PSNL$ datasets. Volunteered users made a two-step evaluation on each tweet for each dataset. They are asked to mark the tweet as interesting or uninteresting, and then if the tweet is interesting, they are asked to score the tweet in the range of $[1-3]$ where 1 is the least score, and 3 is the highest score for interestingness. In the

		Classification Acc. (%)		Ranking Acc. (nDCG)	
		$GNRL$	$PSNL$	$GNRL$	$PSNL$
Inactive Users	$User_1$	47	49	0.520	0.612
	$User_2$	42	39	0.573	0.654
	$User_3$	36	37	0.433	0.478
	$User_4$	43	36	0.322	0.301
	$User_5$	49	47	0.567	0.514
Average (IU)		**43.40**	**41.60**	**0.483**	**0.512**
Active Users	$User_6$	68	64	0.777	0.909
	$User_7$	66	61	0.699	0.768
	$User_8$	62	56	0.760	0.782
	$User_9$	71	72	0.720	0.815
	$User_{10}$	72	65	0.601	0.677
Average (AU)		**67.80**	**63.60**	**0.711**	**0.790**
Average (Overall)		**54.10**		**0.624**	

Table 1: Tweet Recommendation Experiment Results with respect to the Baseline Method

		Classification Acc. (%)		Ranking Acc. (nDCG)	
		$GNRL$	$PSNL$	$GNRL$	$PSNL$
Inactive Users	$User_1$	69	66	0.723	0.773
	$User_2$	62	58	0.684	0.796
	$User_3$	52	55	0.656	0.616
	$User_4$	67	52	0.590	0.623
	$User_5$	72	69	0.734	0.691
Average (IU)		**64.40**	**60.00**	**0.677**	**0.700**
Active Users	$User_6$	88	86	0.809	0.958
	$User_7$	79	74	0.795	0.888
	$User_8$	74	68	0.812	0.826
	$User_9$	88	85	0.815	0.904
	$User_{10}$	80	77	0.773	0.872
Average (AU)		**81.80**	**78**	**0.801**	**0.890**
Average (Overall)		**71.05**		**0.767**	

Table 2: Tweet Recommendation Experiment Results with Respect to the Proposed Method

baseline method, followee rankings are neglected and hence every named entity has equal weight. Generated recommendations are compared against the user preferences in terms of classification, and ranking.

The results in Table 1 show that the baseline method is able to decide whether a tweet is interesting for a user or not with the accuracy of 54,10% on average with classification and 0,624 *nDCG* value on average with ranking, which are lower than the results of our system. The performance of the baseline method in some cases decreases down to 36% correct prediction at classification, and 0,322 *nDCG* value at ranking quality. On the other hand, the results shown in Table 2 shows that the proposed system is able to decide whether a tweet is interesting for a user or not with the accuracy of 71,05% on average for classification and 0,767 *nDCG* value on average for ranking. Given the suitable user habits, performance of the system increases up to the 88% correct prediction for classification, and 0,958 *nDCG* value at ranking quality. The comparison of two tables show that the proposed user interest modeling approach increases the performance.

4. CONCLUSIONS

This paper proposes a new approach to Twitter user modeling and tweet recommendation by making use of named entities extracted from tweets. A powerful aspect of NER approach adopted in this study, tweet segmentation, is that it does not require an annotated large volume of training data to extract named entities, therefore a huge overload of annotation is avoided. In addition, this approach is not dependent on the morphology of the language. Experimental results show that the proposed method is capable of deciding on tweets to be recommended according to the user's interest. Experimental results show the applicability of the approach for recommending tweets.

5. REFERENCES

[1] D. Downey, M. Broadhead, and O. Etzioni. Locating complex named entities in web text. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, IJCAI'07, pages 2733–2739, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[2] V. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

[3] C. Li, A. Sun, J. Weng, and Q. He. Exploiting hybrid contexts for tweet segmentation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 523–532, New York, NY, USA, 2013. ACM.

[4] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee. Twiner: named entity recognition in targeted twitter stream. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 721–730, New York, NY, USA, 2012. ACM.

[5] T. Sezer. TS Corpus, The Turkish Corpus, 2014. [Online; accessed 14-December-2014].

[6] Twitter. About twitter, inc., 2014. [Online; accessed 14-December-2014].

Section II:

Posters & Extended Abstracts

Connections between Twitter Spammer Categories

Gordon Edwards
School of Informatics
University of Edinburgh
Edinburgh, Scotland, UK
g.n.edwards@sms.ed.ac.uk

Amy Guy
School of Informatics
University of Edinburgh
Edinburgh, Scotland, UK
Amy.Guy@ed.ac.uk

ABSTRACT

Twitter has become a viable platform for spammers, who often form networks to further their reach. Troublesomely, targeted users become increasingly frustrated, or worse, view content resulting in computer virus infection. We build on previous work around detecting spam on Twitter, proposing that subcategorising spammers can increase our understanding of their connections in spammer networks and aid detection. After defining five subcategories of spammers and classifying users accordingly, correlations between the categories of spammers and the categories of their followers and followees are explored. We also find that all spam subcategories follow a higher share of non-spam accounts than any individual spam subcategories, and, unexpectedly, that every spammer subcategory is followed by non-spammers more than by individual counterparts.

Keywords

Twitter, spammer categories, spam, social media, microposts, machine learning

1. INTRODUCTION

Twitter's popularity attracts spammers, providing them with a very publicly-accessible user base. It reported that less than 5% of its users are spammers, but that figure is likely to be higher in reality [2], especially with the more wide-ranging criteria for spam adopted in this paper. Spam can pose a security threat to users, or just cause annoyance — either way leaving them disillusioned with Twitter.

Users are not compelled to follow accounts they deem to be spam. However, the ability to quickly determine if a new follower is a spammer is useful in deciding whether to follow back. Automatic detection could save users from wasting time checking each new follower, and spare them from potentially dangerous spam. Spammers can also reach users via a mention or a direct message; in this case investigating the tweet author safeguards against spam.

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

It is suggested in [7] that spammers collude within Twitter networks — that if each account is a node in a graph, then from each node a spam account can be reached by traversing five edges with probability $p = 0.63$. Working together in networks helps spammers proliferate, as it is unlikely a whole network will successfully be taken down. Adding new accounts to their network as others are removed, each can rely on follows from accounts within the network. A desirable but false impression of popularity is thus given. Detecting and classifying whole spammer networks at once could enable more efficient elimination of spam, compared to assessing on a continual basis all individual accounts on the site.

Previous work considers various machine learning techniques for detecting spam, such as Random Forest and Naïve Bayes, either from live feeds or from research corpora [1, 4]. Broadly, it refers to two sets of features upon which users can be classified: content-based, such as mean number of hashtags per tweet, and user-based, such as number of followers of the authoring user [4].

The preceding literature frames spam classification as a binary process (not spam/spam). However, further investigation reveals recurring subtypes of spam—for example users advertising products, or users disseminating pornography—providing a novel approach to classification. Aside from academic interest, classifying into subtypes means users could engage in more refined decisions about blocking of content or users than Twitter's spam filtering currently allows. It also facilitates pinpointing of the most harmful spam, such as tweets concealing viruses and phishing attacks.

Emergent trends, which we will examine, in the distribution of an account's followers and those they follow between the categories may increase confidence that it belongs to a particular category. Finding that one spammer is commonly connected to a particular type yields a fast way to discover accounts of that type, potentially to block or suspend. Connections between different spammer categories are not very dangerous in themselves—though could lure a user to viewing further spam accounts—but they form a potential means of detecting spammer networks.

This paper, part of an ongoing research project, lays the groundwork for investigating the extent to which different categories of spammers are connected to others, and to genuine users. It establishes that these connections result from

spammers' collusion within networks. We build on the work of [7], but contrastingly not confining ourselves to just one trending topic. In Section 2 we describe our defined subcategories of spam, training set, features, and classifier. We then summarise our findings in Section 3 and their limitations in Section 4.

2. CLASSIFICATION

2.1 Spam Subcategories

The Twitter API [6] offers the means to collect a sample of 1,420 users to form a training set, to subsequently hand-label as *spam* and *not spam*. During this annotation process spam subcategories become apparent. Whilst not necessarily definitive, they are reasonably defensible. Though applicable to users and tweets, we only use the categories in relation to users. They are defined below with example tweets typical from the type of spammer. Their distribution is displayed in Figure 1.

- *advertising*: users who tweet extremely frequently, mostly, if not always, advertising products, or tweets advertising a product authored by such a user. Normally the tweets contain links, often shortened using a URL shortener.

- *explicit*: users who post exclusively, or almost so, photos, videos, and links, perhaps shortened with a URL shortener, to websites of a pornographic or adult nature, or tweets that that contain this kind of content.

- *follower gain*: users claiming the ability to boost other users' follower bases, frequently, in most of their tweets, asking users for retweets and to follow certain accounts. A tweet in this category claims that retweeting or following a mentioned (via *@username*) account will result in the receipt of followers.

- *celebrity*: users who tweet plead relentlessly for the follow back of a public figure in their tweets. Ascertaining whether an individual tweet falls into this category is generally harder. Examining the authoring user should be indicative — ascertaining whether a suspect tweet is a unique occurrence for that user and therefore not representative.

- *bot*: accounts whose tweets are generated by a bot that auto-posts content from some source, or details usage of an online app. Tweets that fall into this category often contain a URL, but, again, to be certain in classification the authoring account may need to be examined.

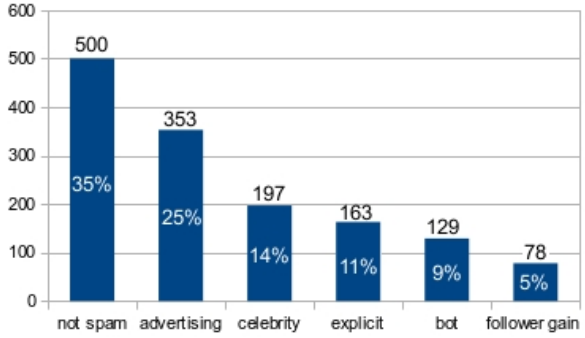


Figure 1: Class distribution of the dataset

2.2 Features

Feature representations of Twitter users can be formed, as per previous work, using content-based and user-based features [4]. Fifty features, 15 user-based and 35 content-based, sufficiently represent users. The content features require the tweet history of the user: their latest 200 tweets, or fewer if they do not have that many. Some features unique to this paper are:

User-based Features
Screen name and description Levenshtein similarity[1]
Percentage of non-alphanumeric characters in description

Content-based Features
Mean number of new lines in the user's tweets
Relative standard deviation of the number of new lines in the user's tweets

2.3 Classifier

The `Random Forest` classifier implementation in the `Weka` Java library [5] provides the basis for implementing a classifier tailored to the spam subcategory classification task. Maximising the spam recall desirably increases the probability of classifying a spammer's spam followers and followees[2] into the subcategories correctly. Thus, the classifier first binarily classifies users as *not spam* and *spam*, using the `Random Forest` classifier — considering all instances labelled as one of the spam subcategories as labelled *spam*. Then, if the outputted classification is *not spam* and the associated confidence is not less than a set threshold[3], *not spam* is returned. Otherwise, the instance is reclassified, again with

[1]Description of Levenshtein similarity: `www.cs.tufts.edu/comp/150GEN/classpages/Levenshtein.html`

[2]For the purposes of this paper "followees" refer to the accounts which a user is following.

[3]Given threshold α, instances initially classified with the binary classifier *not spam*, with confidence c, $c \leq \alpha$, are

the `Random Forest` classifier, applied to dataset with the *not spam* instances filtered out, so one of the spam subcategories is necessarily returned. Conveniently, using Weka's `AdaBoostM1` implementation furthers reduces misclassification due to class imbalance.

Ten-fold cross-validation, provided through `Weka`, allows the classifier to be evaluated, with the collected sample of 1,420 users forming the validation set:

	Recall	Precision	F-Measure
not spam	0.74	0.80	0.77
explicit	0.77	0.83	0.80
advertising	0.84	0.64	0.72
follower gain	0.56	0.90	0.69
bot	0.36	0.56	0.44
celebrity	0.78	0.74	0.76

The classifier performs poorly on the class *bot*, most often misclassifying as *advertising*, so there can be no confidence in conclusions made regarding that class. The misclassification is probably due to the inherent similarity between the behaviours of spammers in each category.

2.4 Results Reporting

For each class, given a sample of 70 contained users the tailored classifier can be used to attain the mean class percentages of followers and followees — 500 (or as many as there are) are sampled for each. Given more time and computational resources, a larger dataset could be formed. All the percentages are rounded to the nearest integer.

Contingency tables are also constructed given the counts of *(category, follower category)* pairs and *(category, followee category)* pairs. These help reveal the extent to which spammers are connected to their followers and to their followees.

3. DISCUSSION OF RESULTS

Possible inaccuracies in classifications detailed in Section 4 mean care should be taken in drawing conclusions, and it is unlikely all of them will be infallible. The results report that genuine users have 73% *not spam* followers on average, 20% higher than the *not spam* followers share of *advertising* and *bot* accounts. Tallying with our intuition, the fair conclusion to draw here given the classifier performance on these follower classes for *not spam* is that genuine users will have a noticeably higher share of *not spam* followers than spammers, a trait that can increase the confidence that a user classified as *not spam* is indeed so. With a fair degree of confidence the results show that genuine users are likely to follow back around half of their genuine followers. The reported number of followers and followees for accounts that spammers follow back is usually higher than for accounts they do not, implying that spammers target their connections to popular accounts.

The average share of *not spam* accounts followed across the *advertising*, *bot*, *celebrity*, and *follower gain* categories, 60%, is notably higher than that of any of the spam subcategories, showing their persistent efforts to gain genuine users' attention. However, perhaps surprisingly, on average 50% of

assumed to be *spam*, to further increase the *spam* recall.

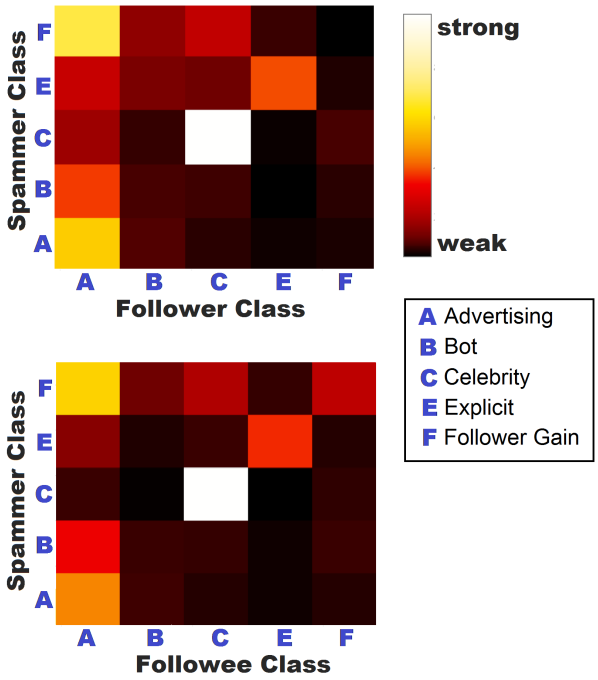


Figure 2: Heat maps showing respectively the strength of connection between spammer subcategories and their follower subcategories, and between spammer subcategories and their followee subcategories.

a spammer's followers are genuine users for each subcategory. Users are either consciously following spammers—perhaps *advertising* accounts hoping to find good deals or *celebrity* accounts because they are interested in the associated celebrity—or through ignorance, lacking a tool to warn them. No one spam category is a landslide winner in attaining genuine followers though.

On average, about 30% of *advertising* followers belong to the same category — a share much higher than any other spam subcategory. Also around 23% of the accounts followed are *advertising*—again, a share much higher than the other spam subcategories—suggesting a significant degree of connection between *advertising* accounts, confirmed later.

Other subcategories appearing to have a high degree of intra-connection are *explicit* and *celebrity*. Accounts in the former have a higher share of *explicit* followers than any other follower subcategory, averaging at 20%, and also follow more accounts of the same subcategory than the others, with a share averaging around 42%. Accounts in the latter have a higher share of *celebrity* followers than the other follower subcategories, averaging at 33%. Such accounts also follow more accounts of the same subcategory than the others, with a share averaging around 42%.

However, accounts in the *bot* category have a higher share, averaging at 26%, of *advertising* followers than *bot* followers (averaging at only 6%) or any other subcategory of follower. Likewise the followees share is higher for *advertising*, averaging at 18%, than *bot* (averaging at only 4%) and the other subcategories. This discrepancy could be due to the categories' inherent similarity; arguably both have the same

Class	Follower	Recall	Precision	F1
advertising	*advertising*	0.51	0.60	0.59
	bot	0.43	0.74	0.55
	not spam	0.57	0.43	0.49
bot	*advertising*	0.56	0.63	0.59
	bot	0.44	0.55	0.49
	not spam	0.48	0.58	0.52
celebrity	*celebrity*	0.63	0.50	0.56
	not spam	0.37	0.93	0.53
explicit	*explicit*	0.43	1.0	0.6
follower gain	*not spam*	0.39	0.83	0.53
not spam	*follower gain*	1.0	0.50	0.67
	not spam	0.44	0.98	0.61

Table 1: For each subcategory of spammer the performance when the classifying each subcategory of follower.

aim— to direct users to content—so there is incentive for them to connect with each other. As previously warned, given the categories are not definitive, *advertising* and *bot* could reasonably be merged into one category, probably reducing the classification error.

We confirm the hypothesised relationships in the connections between spammers of the same subcategory using Cramér's V correlation ϕ_c [3]. Measuring the correlation between two categorical random variables given a constructed contingency table, it ranges from 0, where the two random variables are independent, to 1, where they are equal. Letting `X = Subcategory of spammer` and Y = `Subcategory of follower`, $\phi_c = 0.39$, showing that there is some association between a spammer subcategory and their follower subcategory. Similarly, if `Subcategory of spammer` and Y = `Subcategory of followee`, then $\phi_c = 0.47$, showing there is an analogous correlation between a spammer subcategory and their followee subcategory.

The fairly strong positive correlations and attained percentage shares aforementioned evidence the degree of collusion between spammers, and that those in the same subcategories are deliberately connecting to form networks — notable relationships are present. Predicated on these correlations, the heat maps in Figure 2 show the strength of spammer connections. Because it is a hallmark of spam, establishing the presence of such connections aids spammer network detection and individual account classification.

4. LIMITATIONS

When the classifier is further tested by classifying a sample of followers of users from each of the categories, the performance reported in Table 4 is worse than the cross-validation in Section 2.3, likely due to large variations in the distribution as the sample is more deterministic than the validation set. Thus in Section 3 only sound conclusions respecting these figures were drawn, but improvements made in future work could allow further conclusions regarding the connections between some of the combinations of categories not considered. A larger test sample, perhaps yielding different figures, would clearly be preferable but was not practicable given the time constraints.

5. SUMMARY AND FUTURE WORK

This paper presents the findings of new research. By forming a training set of users and implementing a classifier tailored to the task, underpinned by Random Forest, users can be classified into the defined classes. Analysing the distribution of these classes in users' followers and followees allows inferences to be made about the relationships between users, crucially between spammers. We observe that many genuine users are falling into the trap of connecting with a range of types of spammer.

We reveal that spammers mainly have their largest share of connections devoted to non-spammers and their second largest to spammers of the same subcategory. However there are exceptions, with some subcategories connecting with a proportionally very much smaller number of spammers from the same category. Correlations are found between spammer subcategories and their follower and followee subcategories, showing that spammers are colluding with each other in networks, with a significant degree of connection between spammers of the same category.

Establishing connections between subcategories in a large contiguous network, starting from one account and branching outwards, recursively analysing the followers and followees, could be a future extension. Visualising this network would be interesting, allowing clusters of spammers of different subcategories to be determined. Also the subcategories could usefully be refined, and perhaps more introduced.

6. ACKNOWLEDGMENTS

We thank Krzysztof Jerzy Geras, School of Informatics, University of Edinburgh, for explaining to us how to find correlations, which we subsequently found and included in this paper.

7. REFERENCES

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.

[2] J. Brustein. Twitter's bot census didn't actually happen. `http://www.businessweek.com/articles/2014-08-12/twitters-bot-population%-remains-a-mystery-and-a-problem`. [Online; accessed 14/11/2014].

[3] P. Dattalo. Nominal association: Phi and cramer's v. `http://www.people.vcu.edu/~pdattalo/702SuppRead/MeasAssoc/NominalAssoc.%html`, 2002. [Online; accessed 10/03/2015].

[4] M. McCord and M. Chuah. Spam detection on twitter using traditional classifiers. In *Proceedings of the 8th International Conference on Autonomic and Trusted Computing*, ATC'11, pages 175–186, Berlin, Heidelberg, 2011. Springer-Verlag.

[5] N. Z. The University of Waikato. Weka. `http://www.cs.waikato.ac.nz/ml/weka/`.

[6] Twitter4j. Twitter4j. `http://twitter4j.org/`.

[7] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd. Detecting spam in a twitter network. In *Volume 15, Number 1 - 4 January 2010, First Monday peer-reviewed journal*. First Monday, 2010.

A Topical Crawler for Uncovering Hidden Communities of Extremist Micro-Bloggers on Tumblr

Swati Agarwal
Indraprastha Institute of Information Technology, Delhi (IIIT-D), India
swatia@iiitd.ac.in

Ashish Sureka
Software Analytics Research Lab (SARL), India
ashish@iiitd.ac.in

ABSTRACT

Research shows that microblogging websites such as Tumblr are being misused as a platform to disseminate hate and extremism. We formulate the problem of locating such extremist communities as a graph search problem. We propose a topical crawler based approach performing several tasks: searching for a blogger, computing its similarity against exemplary documents, filtering hate promoting bloggers, navigating through links to other bloggers and managing a queue of such bloggers for social network analysis. We conduct experiments on real world dataset and examine the effectiveness of 'like' and 'reblog' features as links between bloggers. Experimental results demonstrates that the proposed solution approach is effective with an F-score of 0.80.

Keywords

Mining User Generated Content, Online Radicalization, Social Media Analytics

1. PROBLEM DEFINITION & SOLUTION

Tumblr is a popular and widely-used micro-blogging website. Previous research shows that such websites are used as a platform for disseminating hate and extremism (due to low barrier to publication and anonymity) [1][2][3][4][5]. Automatic identification of hate and extremism promoting posts and bloggers is an important (from the perspective of the website moderators and law enforcement agencies) and a technically challenging problem. Large volume of data on Tumblr, free-form text and noisy content makes automated analysis technically challenging [1][2][3][4][5]. Our aim is to investigate the application of a topical crawling based algorithm for retrieving hate promoting bloggers on Tumblr. Our objective is to examine the effectiveness of a random-walk based approach in social network graph traversal. Furthermore, our goal is to examine the effectiveness of *re-blogging* and *like* on a post as the links between two bloggers and conduct experiments on large real world dataset to demonstrate the effectiveness of our approach.

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

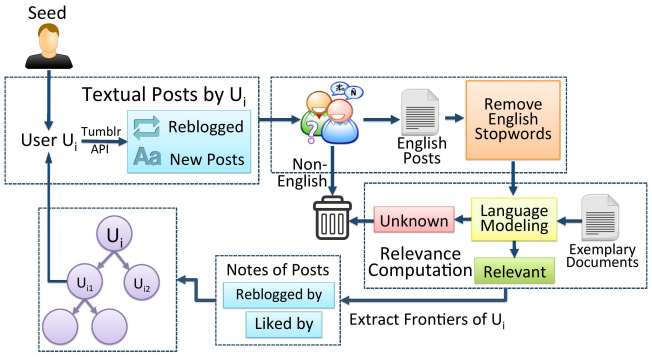


Figure 1: Proposed Architecture for Extremist Community Detection

In a graph traversal, a topical crawler returns relevant nodes to a specific topic. To define the relevance of a node, it learns the characteristics and features of given topic and computes the extent of similarity against a bunch of exemplary documents. To collect training examples, we perform an iterative search on Tumblr using keyword based flagging, where keyword is a search tag; for example, jihad, anti-Islam and hate. We perform a case study on Jihad and by manual search on Tumblr posts we collect several relevant tags that are commonly used by extremist bloggers. We use these tags to initiate our process and collect all textual posts (avoiding picture, audio, video and URLs), tags (associated with resultant posts) and linked bloggers (post reblogged by and liked by) with no redundancy. We perform a manual inspection on resultant posts and posts made by linked bloggers to filter relevant (hate promoting) and unknown results. We further extract more posts and linked bloggers from related tags and run this framework recursively to collect our exemplary documents (400 hate promoting posts). These training examples contain the body and caption of only positive class (hate and extremism promoting content) posts which is used to train the model.

Figure 1 illustrates the design and architecture of topical crawler to locate extremist communities. As shown in Figure 1, our proposed solution framework is an iterative multi-step process primarily consisting of five phases: features (posts) extraction, data pre-processing, classification, frontier extraction and graph traversal. In phase 1, we initiate our process using a positive class (hate promoting) blogger U_i called

as 'seed'. We use Tumblr API [1] to fetch the URLs of n number of textual posts and by using Jsoup Java library [2] we extract the content and caption of these posts (used as contextual metadata). These posts can be either re-blogged from other users or originally posted by the user U_i. These posts consist of multiple langauges. Therefore, in phase 2, we perform data pre-processing and filter English and non-English posts using language detection library[3]. We perform data pre-processing on these posts and remove English stopwords. In phase 3, we build a statistical model from the exemplary documents collected separately by semi-automatic process. To compute the relevance of each blogger, we use character level n-gram language modeling approach. We find the extent of similarity between metadata and exemplary documents using LingPipe API [4] - applying joint probability-based classification of character sequences . We implement a one class classifier and filter extremism promoting bloggers from unknown bloggers. In phase 4, we extract the notes associated with the posts (collected in phase 1) of relevant bloggers. These notes contain the list of bloggers who liked and re-blogged a particular post. The number of notes represent the popularity of a post and indicate the similar interest between original poster and other bloggers in the list who may or may not be the direct followers of each other. We use notes to extract frontier nodes of a blogger because of two reasons: 1) due to the privacy policies Tumblr API does not allow developers to extract followers and following blogs of Tumblr users. 2) Tumblr facilitates bloggers to track any number of tags so that whenever there is a new post published publicly on Tumblr containing any of these tags, it automatically appears in a menu on user's dashboard. They can spread that post among their followers by re-blogging it. Tracked tags allow bloggers to form a virtual community without following each other. For each frontier extracted in phase 4, we compute the relevance score against exemplary documents and discard unknown bloggers. In phase 5, we manage a queue of relevant bloggers and perform directed graph traversal using random walk algorithm. To expand our graph we select the next blogger in uniform distribution and extract it's frontiers. We execute our focused crawler for each frontier without revisiting a blogger. This traversal results in a connected graph, where nodes represents a blogger (hate promoting) and edges represent the links (re-blog and like) between two bloggers. We perform social network analysis on the resultant graph and locate extreme right communities of hate promoting bloggers.

2. RESULTS & CONCLUSION

We execute our topical crawler for a given seed blogger and traverse through Tumblr network using random walk algorithm. For every new blogger, we compute its relevance and classify it as hate promoting or unknown using one class classifier. To examine the effectiveness of our classifier, we compute its accuracy using standard information retrieval techniques. In one execution of our topical crawler, we were able to collect 600 bloggers. We hired 30 graduate students as volunteers from different department to label these bloggers as hate promoting or unknown according to their published posts and given guidelines for annotation. To avoid the biasness and to collect correct annotated results we perform a horizontal and vertical partition on nodes and arrange these 600 bloggers into a 2D matrix where rows are the numbers of annotators grouped in 10 sets, 3 members each. Columns of the matrix are the number of bloggers assigned to each member for annotation i.e. 60. We use majority voting approach for final annotation, the class of a blogger is the one which is voted by at least two annotators. Based upon the validation results we evaluate the accuracy of our model. Table 1(a) shows the confusion matrix for one class classification. Table 1(a) reveals that our model predicts 382 (290+92) bloggers as hate promoting and 218 (173+45) bloggers as unknown. Table 1(a) shows that there is a missclassification of 13% and 34% in predicting hate promoting and unknown bloggers. Table 1(b) shows the accuracy results of our classifier. Results shows that the precision, recall and f-score are reasonably high and we are able to predict hate promoting bloggers with an accuracy of 77%. Our experimental analysis reveals that re-blogging is a good indicator of connection between two bloggers. We locate users who are central and influential among all and play major role in the discovered communities. We perform independent social network analysis on like and re-blog links among bloggers and conclude that re-blogging is a discriminatory feature to identify the communities of extremist bloggers sharing a common agenda.

Table 1: Confusion Matrix and Accuracy Results for One Class Classifier

(a) Confusion Matrix

		Predicted	
		Positive	**Unknown**
Actual	**Positive**	290	45
	Unknown	92	173

(b) Accuracy Results

Precision	Recall	F-Score	Accuracy
0.75	0.86	0.80	0.77

[1]https://www.tumblr.com/docs/en/api/v2
[2]http://jsoup.org/apidocs/
[3]https://code.google.com/p/language-detection/
[4]http://alias-i.com/lingpipe/index.html

3. REFERENCES

[1] S. Agarwal and A. Sureka. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *Distributed Computing and Internet Technology (ICDCIT)*, pages 431–442, 2015.

[2] E. A. Cano Basave, Y. He, K. Liu, and J. Zhao. A weakly supervised bayesian model for violence detection in social media. In *Sixth International Joint Conference on Natural Language Processing*, pages 109–117, 2013.

[3] S. Kumar, F. Morstatter, R. Zafarani, and H. Liu. Whom should i follow?: Identifying relevant users during crises. In *ACM Conference on Hypertext and Social Media (HT)*, pages 139–147, 2013.

[4] A. Sureka and S. Agarwal. Learning to classify hate and extremism promoting tweets. In *Joint Conference in Intelligence Security Informatics (JISIC)*, pages 320–320. IEEE, 2014.

[5] J. Xu, T.-C. Lu, R. Compton, and D. Allen. Civil unrest prediction: A tumblr-based exploration. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 403–411. Springer, 2014.

Section III:

Social Sciences Research Track

edited by

Katrin Weller & Danica Radovanović

Matthew Rowe, Milan Stankovic & Aba-Sah Dadzie

Making Sense of Microposts (#Microposts2015) Social Sciences Track

Danica Radovanović*
Faculty of Technical Sciences,
University of Novi Sad, Serbia
danica@danicar.org

Katrin Weller*
GESIS Leibniz Institute for the
Social Sciences, Germany
katrin.weller@gesis.org

Aba-Sah Dadzie*
Knowledge Media Institute,
The Open University, UK
aba-sah.dadzie@open.ac.uk

ABSTRACT

For the first time in its five year history the #Microposts workshop features a designated Social Science track. This paper introduces this new track by situating it within the overall workshop objectives. It highlights the importance of interdisciplinary studies in the attempt to make sense of Web user activities in general, and in the generation and consumption of Microposts in particular. This paper provides examples of related work in the field, such as Computational Social Science, reviews previous contributions to the #Microposts by the Social Science research community, and introduces the two papers presented in the track.

Keywords

Microposts, Social Science, Web Science, Computational Social Science, Internet science, social media, user-generated content, online communication, Internet research

1. INTRODUCTION

The Internet is not just a static set of tools or affordances for a specific set of user-defined purposes. Rather, it also represents a rapidly evolving set of ways to configure one's social life. That is to say, the Internet today enables different relationships within the basic dimensions of social and cultural dynamics and organisation [4]. New media and technology denote embodiments of socio-cultural relationships that in turn shape and structure our possibilities for social action, education and cultural expression [1, 6] across all generations and walks of life. The myriad ways that social lives can be (re-)arranged through various types of media and communication forms however present a challenge for researchers from multiple disciplines.

It can be postulated that social dynamics facilitate new forms of communication structures in social lives. One of those structures present *Microposts* – each a small, brief message, theme or a single thought, quick and easy to publish, and that, posted from a variety of platforms and by very large numbers of individuals with as many viewpoints and interests, collectively provide a rich source of information and opinion about a range of topics. Microposts present a dominant forum in social networks, micro-blogging services and virtual communities, and have become of socio-technological value. In recognition of this, the #Microposts workshop was born, to provide an avenue for different disciplines to come together to *make sense of Microposts*, to identify why they have become and remain a significant means of communication, how the phenomenon impacts its users and the wider society, and how end users today, both the technology-rich and those digitally disadvantaged, make use of the platform and consume the rich content generated in their social and working lives.

*All authors made equal contributions

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

2. THE SOCIAL SCIENCES IN THE ANALYSIS OF MICROPOSTS

Recent years have brought about an increasing number of interdisciplinary approaches, between computer science and social sciences, often also referred to as *Computational Social Science* [8]. Computational Social Science uses computational methods to study social behaviour, e.g., by developing computational approaches that consider empirical methods and theories from social sciences, and by exploring new kinds of data to learn about social phenomena [19]. Different workshops and events are currently being organised in order to discuss new approaches in the field of computational social science and exchange useful approaches as well as experience with new datasets. These include the *International Conference on Computational Social Science in Helsinki*[1], to be held in June 2015. The importance of these connections across the disciplines are now recognised widely; interestingly, while the NEEL (**N**amed **E**ntity r**E**cognition and **L**inking) Challenge[2], which forms part of the #Microposts workshop, typically attracts a select group, due to its specific focus, a social sciences researcher in 2015 tweeted from the WWW'2015 conference: "*An effective named entity recognition for Twitter would be invaluable for social scientists too. Go NEEL #Microposts2015 guys!*"[3].

In trying to make sense of Microposts, researchers may ex-

[1] http://www.iccss2015.eu
[2] http://www.scc.lancs.ac.uk/microposts2015/challenge
[3] Fabio Giglietto [fabiogiglietto] (1:42 PM – 18 May 2015 Tweet) Retrieved from http://bit.ly/1FqJTA0

plore and apply a variety of approaches. The proceedings of the previous #Microposts workshops prove this, as they already include contributions from various academic backgrounds, such as computer science, social sciences, sociology, digital ethnography, psychology and linguistics. In 2013, for instance, Vanin *et al.,* [21] in *Some Clues on Irony Detection in Tweets*, presented a mixed methods study to counter a challenge in automated analysis – interpretation of the particular context, including tweeter style or personality, and even subtleties unique to specific languages. In 2012, Radovanović & Ragnedda [12] presented a study on *Small Talk in the Digital Age: Making Sense of Phatic Posts*, in which they discussed the role of Microposts in social, dynamic communication on the Web, and the value in this medium for end users, in terms of content and for driving the conversation itself. In 2011, the first year in which the workshop was held, Šķilters *et al.,* [16] in *The Pragmatics of Political Messages in Twitter Communication*, carry out detailed content analysis of the participants in the 2010 Latvian parliamentary elections, to identify pragmatic patterns in political communication, based on the identities of individuals and (virtual) communities. In this first workshop, also, Weller *et al.,* [22] in *Citation Analysis in Twitter: Approaches for Defining and Measuring Information Flows within Tweets during Scientific Conferences*, examine a number of features in information exchanged on Twitter during scientific conferences, to provide, within webometrics, an alternative source of citations.

It has always been an aim of the workshop series to bring together computer scientists and researchers from other disciplines, including social scientists. For this reason, we have also sought to include guest speakers with work spanning Computer Science and Social Sciences, including that by Greg Ver Steeg [17] on *Information Theoretic Tools for Social Media* in 2012, Daniele Quercia [11] on *Urban*: Crowdsourcing for the good of London*] in 2013, and Markus Strohmaier [18] on *Computational Social Science and Microblogs – The Good, the Bad and the Ugly* in 2014. To highlight even further this objective, the *#Microposts2015 workshop* [23] features an explicit social sciences track in addition to the main track. By including this track and publishing a specific call for papers for social scientists, we were able to recognise the different publication practices that are one of the current challenges for successfully bringing together researchers from different disciplines.

2.1 Track Sponsor: GESIS

User-generated content and social media data are one major source in computational social science. For example, Microposts from social media platforms can provide new insights into political communication around elections [10, 7], political activism [9, 20] or disaster response [2]. *GESIS*, the Leibniz Institute for the Social Sciences [24], is a research infrastructure and service provider for the social sciences. GESIS hosts one of the first departments in Computational Social Science in Germany, where interdisciplinary researchers develop algorithms and theories for studying social phenomena based on Web data and also organise workshops and training opportunities. As part of the engagement in supporting social scientists in this new field, GESIS is also sponsoring the prize for the best social science paper at the 2015 #Microposts workshop.

3. THE #MICROPOSTS2015 SOCIAL SCIENCES TRACK

For the first dedicated Social Science track in the #Microposts series, three submissions were received, with an additional two from the main track crossing the boundary between this and the main track. Of these, two papers out of the first three were accepted for presentation for the track.

The award for best submission went to the paper *To Be or Not to Be Charlie: Twitter Hashtags as a Discourse and Counter-discourse in the Aftermath of the 2015 Charlie Hebdo Shooting in France* by Giglietto & Lee [5]. Written in the wake of the shooting in Paris, this paper provides one of the first studies of Twitter users' reactions to the event, and examines the human reaction on Twitter, expressing solidarity with the victims in different ways. The analysis examined the viewpoint of tweeters who appeared to oppose what was considered the norm as an expression of solidarity, in how they chose to express their grief and sympathy, and also resistance, using an expression that reinforced their identity with *#JeNeSuisPasCharlie*, in contrast to the spontaneously derived *#JeSuisCharlie* hashtag.

Coelho, Lapa, Ramos & Malini [3] in *A Research Design for the Analysis of Contemporary Social Movements*, looked at political, social empowerment in today's digital culture, through discursive analysis of Microposts. An important contribution of their qualitative study is to help to develop guidelines for teachers, to enable effective, critical appropriation of the data generated on social networks by net activist groups. The aim is to support education of young people, to encourage participation in the social freedom and the socio-political agenda.

Other papers addressing civil and political activism, and the analysis of data generated as a result, due to citizen empowerment and social cohesion, or, in contrast, diversive political activity, were submitted to both the social sciences and the main track. The call for papers highlighted other key topics, some of which also overlapped with the call for the main track. These included data journalism, collective awareness, citizen empowerment and education, and psychological aspects of Micropost-based interactions. Additional topics of particular importance to social science research include inequality in access to and the use of digital media, and how Micropost-based services have resulted in the emergence of alternative social and communication dynamics. The perspectives taken and the approach to data analysis clearly differed from the main track, with the social sciences track focusing not just on data content, but also on the human element that influences the publishing of Microposts, and how its content may be subsequently appropriated in the modern, digital world. We believe the overlap and divergence in approaches reinforces the need for the two fields, along with other relevant disciplines, to work in tandem in the analysis of Micropost data, allowing the different lenses through which each field works to result in increasingly richer analysis of this very diverse and constantly growing data set.

Acknowledgments

Danica Radovanović is an Internet researcher, who graduated from the University of Novi Sad, after research during

her PhD as a Chevening Scholar at the Oxford Internet Institute. Katrin Weller works at the GESIS Leibniz Institute for the Social Sciences and is currently funded by the John W. Kluge Center at the Library of Congress through a fellowship in Digital Studies. Aba-Sah Dadzie is a visiting researcher at KMi, the Open University, and is working on the EU project EDSA (no. 643937).

4. REFERENCES

[1] W. Bijker, T. Hughes, and T. Pinch. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT Press, 1987.

[2] A. Bruns and J. Burgess. Crisis communication in natural disasters: The Queensland floods and Christchurch earthquakes. *Twitter and Society*, pages 373–384, 2014.

[3] I. C. Coelho, A. Lapa, V. Ramos, and F. Malini. A research design for the analysis of contemporary social movements. In Rowe et al. [14], pages –.

[4] W. H. Dutton and G. Blank. Cultures of the internet: Five clusters of attitudes and beliefs among users in Britain. Technical report, OII Working Paper, Oxford Internet Surveys (OxIS) Project, February 2014.

[5] F. Giglietto and Y. Lee. To Be or Not to Be Charlie: Twitter hashtags as a discourse and counter-discourse in the aftermath of the 2015 Charlie Hebdo shooting in France. In Rowe et al. [14], pages –.

[6] M. Ito. *Hanging Out, Messing Around, Geeking Out: Living and Learning with New Media*. MIT Press, 2009.

[7] A. Jungherr, P. Jürgens, and H. Schoen. Why the Pirate Party won the German election of 2009 or the trouble with predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. 'Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment'. *Social Science Computer Review*, 30(2):229–234, 2012.

[8] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Computational Social Science. *Science*, 323(5915):721–723, 2009.

[9] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and d. boyd. The Arab Spring – the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, 5(0), 2011.

[10] H. Moe and A. O. Larsson. Untangling a complex media system. *Information, Communication & Society*, 16(5):775–794, 2013.

[11] D. Quercia. Urban: Crowdsourcing for the good of London. In *Proc., 22nd International Conference on World Wide Web (WWW '13 Companion)*, pages 591–592, 2013.

[12] D. Radovanović and M. Ragnedda. Small talk in the digital age: Making sense of phatic posts. In Rowe et al. [13], pages 10–13.

[13] M. Rowe, M. Stankovic, and A.-S. Dadzie, editors. *Proceedings, 2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages, Lyon, France, 16 April 2012*, April 2012.

[14] M. Rowe, M. Stankovic, and A.-S. Dadzie, editors. *Proceedings, 5th Workshop on Making Sense of Microposts (#Microposts2015): Big things come in small packages, Florence, Italy, 18th of May 2015*, May 2015.

[15] M. Rowe, M. Stankovic, A.-S. Dadzie, and M. Hardey, editors. *Proceedings, 1st Workshop on Making Sense of Microposts (#MSM2011): Big things come in small packages, Heraklion, Crete, Greece, 30th May 2011*, May 2011.

[16] J. Šķilters, M. Kreile, U. Bojārs, I. Brikše, J. Pencis, and L. Uzule. The pragmatics of political messages in Twitter communication. In Rowe et al. [15], pages 69–80.

[17] G. V. Steeg. Information theoretic tools for social media. In Rowe et al. [13], pages 1–1.

[18] M. Strohmaier. Computational Social Science and microblogs – the good, the bad and the ugly. In M. Rowe, M. Stankovic, and A.-S. Dadzie, editors, *4th Workshop on Making Sense of Microposts (#Microposts2014)*, pages 1–1, April 2014.

[19] M. Strohmaier and C. Wagner. Computational Social Science for the World Wide Web. *IEEE Intelligent Systems*, 29(5):84–88, Sept 2014.

[20] K. Thorson, K. Driscoll, B. Ekdale, S. Edgerly, L. G. Thompson, A. Schrock, L. Swartz, E. K. Vraga, and C. Wells. Youtube, Twitter and the Occupy Movement. *Information, Communication & Society*, 16(3):421–451, 2013.

[21] A. A. Vanin, L. A. Freitas, R. Vieira, and M. Bochernitsan. Some clues on irony detection in tweets. In *Proc., 22nd International Conference on World Wide Web (WWW '13 Companion)*, pages 635–636, 2013.

[22] K. Weller, E. Dröge, and C. Puschmann. Citation analysis in Twitter: Approaches for defining and measuring information flows within tweets during scientific conferences. In Rowe et al. [15], pages 1–12.

[23] #Microposts2015 website. `http://www.scc.lancs.ac.uk/microposts2015`.

[24] GESIS. `http://www.gesis.org`.

Section IIIa:

Social Sciences Submissions

To Be or Not to Be Charlie: Twitter Hashtags as a Discourse and Counter-discourse in the Aftermath of the 2015 Charlie Hebdo Shooting in France

Fabio Giglietto
DISCUM, Università di Urbino Carlo Bo
ITALY
fabio.giglietto@uniurb.it

Yenn Lee
SOAS, University of London
UK
yl22@soas.ac.uk

ABSTRACT
Following a shooting attack by two self-proclaimed Islamist gunmen at the offices of French satirical weekly *Charlie Hebdo* on 7th January 2015, there emerged the hashtag #JeSuisCharlie on Twitter as an expression of condolences for the victims, solidarity, and support for the magazine's right to free speech. Almost simultaneously, however, there was also #JeNeSuisPasCharlie explicitly countering the former, affirmative hashtag. In this paper, we analyse 74,047 tweets containing #JeNeSuisPasCharlie posted between 7th and 11th January. Our network analysis and semantic cluster analysis of those 74,047 tweets reveal that the hashtag in question constituted a form of resistance to the mainstream framing of the issue as freedom of expression being threatened by religious intolerance and violence. The resistance was manifested through three phases: sharing condolences but indicating a reservation against the mainstream frame (Grief); voicing out resistance against the frame (Resistance); and developing and deploying alternative frames such as hate speech, Eurocentrism, and Islamophobia (Alternatives). The hashtag in this context served as a vehicle through which users formed, enhanced, and declared their self-identity.

Categories and Subject Descriptors
J.4 [**Social and behavioral sciences**]: Sociology.

General Terms
Human Factors.

Keywords
counter-discourse, freedom of expression, hashtag, identity, semantic cluster analysis

1.INTRODUCTION
On 7th January 2015, two gunmen forced their way into and opened fire in the headquarters of satirical weekly magazine *Charlie Hebdo* in Paris, killing twelve staff cartoonists and claiming that it was an act of revenge against the magazine's portrayals of the Prophet Mohammed. Within hours following the attack, the hashtag #JeSuisCharlie [I am Charlie] began trending on Twitter, in a show of condolences for the victims, solidarity, and support for the magazine's right to satirise any subject including religions. Reportedly created by an artist named Joachim Roncin, who lived in the neighborhood of the shooting site, the hashtag was used over five million times by 9th January and became one of the most repeated news-related hashtags in Twitter's history [22]. In the initiator's own words, 'je' in this context was important as it offered a vehicle through which each individual expressed themselves vis-à-vis threats to the freedom and tolerance underpinning the participants' world (Roncin, interviewed by Sky News, 2015). 'Je Suis Charlie' (and by extension 'Nous Sommes Tous Charlie' [We are all Charlie]) also served as the principal slogan during the vigils and marches that took place in central Paris on Sunday 11th January.

However, there too emerged #JeNeSuisPasCharlie [I am not Charlie], explicitly countering the former, affirmative hashtag. Since the former hashtag entailed a tragedy of twelve deaths and support for the universal value of freedom of expression, #JeNeSuisPasCharlie carried an inherent risk of being viewed as opposing accepted social norms. Despite the risk, the negative hashtag was used more than 74,000 times over the next few days since 7th January. Against this backdrop, we set out to unpack a complex relationship between the willingness to speak up on sensitive topics and identity formation on Twitter. More specifically, we aim to address three interlinked questions as below.

1. What are the characteristics of the network formed around the #JeNeSuisPasCharlie hashtag and the material shared through that network on Twitter?
2. How did users of the #JeNeSuisPasCharlie hashtag position themselves discursively with regard to the #JeSuisCharlie hashtag?
3. How did the activities under the #JeNeSuisPasCharlie hashtag evolve as the broader public discussion of the shooting attack developed?

2.LITERATURE REVIEW
In order to address the research questions above, the present study draws upon a combination of three strands of work in the current scholarship: the network characteristics of Twitter-mediated discussion; the roles of hashtags in such discussion; and the expressions of identity in social media activism. First, recent years have seen a fast-growing body of literature concerned with buzzing discussions on the microblogging platform Twitter and how to examine them systematically. Given the range and amount of data that researchers could mine from the platform, a keen

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

interest has been shown in employing network-analysis approaches for a 'bird's eye view'. Himelboim and Han [10] argued, through their case study of cancer-related discussion on Twitter, that communities emerged from such discussion with clusters of interconnected users and the information sources on which they relied most. A 2014 special issue of *American Behavioral Scientist*, particularly the contributions by Dubois and Gaffney [5] and Xu et al. [23], showed that opinion leaders and influencers could be metrically identified in Twitter-mediated political discussions. The links formed between political discussants on Twitter turned out to be considerably different from those observed in the Web 1.0 environment or in blogosphere, at least in the South Korean context, according to Hsu and Park [11]. Mapping the landscape of Twitter activity has provided unique insights into various issues of international relevance. Lotan's study of the 2014 Israel-Gaza conflict [12], for example, visually demonstrated a distinct polarisation between the pro-Israel and pro-Palestine sides with a negligible number of bridging actors in-between. By tracing the Twitter network of Western-origin Jihad fighters, Klausen [14] identified that certain strategic roles were assigned to those fighters' Twitter accounts.

Discussions on Twitter are speedy and unstructured and, consequently, the organisational usefulness of hashtags has attracted practical as well as academic attention. Bruns [3] detailed out his methodological experiences and reflections of handling Twitter data around a hashtag and highlighted that hashtags are 'shared conversation markers', which require users to include them in their posts deliberately if they wish to take part in established conversations. Based on a comparison of various hashtag-based communications, Bruns and Stieglitz [4] concluded that different hashtags are associated with different patterns of user behaviours. While crisis- and emergency-related hashtags (such as #tsunami for the March 2011 tsunami in Japan and #londonriots in 2011) have seen a dominant proportion of retweets and URLs pointing outside Twitter, spectacle-oriented hashtags (such as British #royalwedding in 2011 and #eurovision for the Eurovision Song Contest in 2011) seem to elicit more original tweets from users. Indeed, such findings from hashtag studies are in line with the studies focusing on unravelling the network properties of Twitter communications discussed earlier. Siapera's work on #Palestine [19] and Lorentzen's work on #svpol (for Swedish politics) [8], for example, point to homophily and polarisation in hashtag-based discussions, resonating Lotan's findings cited above.

However enthusiastic the participants in Twitter-mediated political discussions may be, whether their participations lead to any concrete outcomes is still an ongoing question. On the one hand, some offer encouraging anecdotes of how Twitter has facilitated protests in different parts of the world, such as one against police brutality in Ferguson in Missouri, US, in 2014 [9]. A cautious voice, on the other hand, is that Twitter and other such platforms make social movements 'easier to organise but harder to win' by pushing them to scale up before they are ready for it [21]. Nevertheless, what social media including Twitter can certainly provide is a space for accommodating expressions of identity at multiple layers. Bennett and Segerberg [2] suggested that, in today's large-scale 'connective action' (in distinction to the traditional concept of 'collective action'), political content is often presented in the form of easily personalised ideas such as 'Put People First' (PPF) during the 2009 G20 London summit protests or 'We Are the 99 Percent' during the Occupy Wall Street movement in the US in 2011. According to the two authors, these personal action frames are particularly inclusive and can be easily passed across different platforms. 'Identity' here can be a collective identity expressed within a limited time span like during one TV programme [1] or a series [7]. More relevantly to the purposes of the present study, identity may refer to individuality that used to be blended and lost in the presence of the collectivity required in activism in the pre-social media era [18].

3.METHODOLOGY

Our dataset consisted of 74,074 tweets containing the hashtag #JeNeSuisPasCharlie and published by 41,687 unique users between 7th and 11th of January 2015. Due to the known limits of Twitter free API [17], the data was purchased from Sifter, a web application that provides, in partnership with Gnip, search-and-retrieve access to every undeleted tweet in the history of Twitter. The data gathered via Sifter was automatically imported into a new DiscoverText project. It was then exported in CSV format from there and was analysed using R.

3.1.Typology of contents and network

The first tweet in the dataset was dated 7th January 2015, 1:46 PM in local time. The hashtag #JeSuisCharlie was reported to be created at 12:59 PM on the same day, immediately following the shooting that took place at around 11:30 AM. Tweets in our dataset were written in various languages. Using the text categorisation engine based on n-grams provided by the textcat R package [6], we discovered that French (30%), English (25%) and Spanish (12%) accounted for the majority of the tweets. It was unsurprising that French was the most frequently used language, but the proportion was smaller than expected, indicating its reference to #JeSuisCharlie. Another interesting characteristic identified was that 1,488 tweets (2%) were made of nothing but the #JeNeSuisPasCharlie hashtag. 70% of the 74,074 tweets were retweets and 41% included URLs. Since retweets account for almost three quarters of the dataset, we computed and visualised a retweet network with a view to identifying central users and their clusters if any. We also identified the most recurring external sources (URLs).

3.2.Topics

In order to understand the main topics addressed, we applied the text mining techniques provided by the textcat R package [16] to the textual corpus of all tweets in the dataset. We lowered the case of all terms in the corpus and cleaned it up by removing auxiliary words in French, English and Spanish, as well as punctuation marks and whitespaces. Additionally, we also removed 'jenesuispascharlie', 'charlie', 'charliehebdo', 'hebdo', 'jesuischarlie' and created a document term matrix to calculate the associations between the remaining words (N=36,030). After removing sparse terms (i.e. the sparsity of a term is defined as the percentage of documents with 0 occurrence; in the present study a term was removed if its sparsity was higher than 98%), we identified the most frequently used terms (N=17) and their Euclidean distances, and created clusters of frequently co-occurring terms.

3.3.Evolution over time

To better understand the evolution of the topics discussed, with particular reference to our third research question, we created a by-minute time series (N=6,444, AVG TPM=11.5) of activity. We also used the Breakout Detection R package, which had recently been open-sourced by Twitter [13], to identify breakouts or shifts in the mean of tweet per minute (TPM).

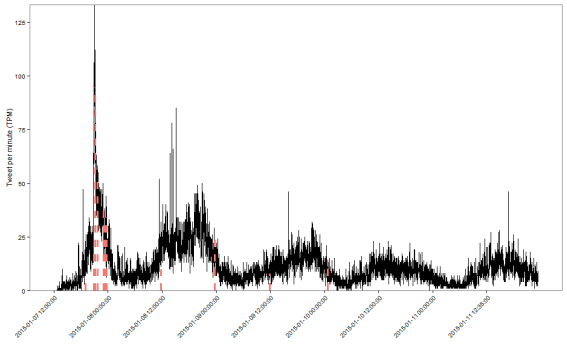


Figure 1. Twitter by-minute activity on the hashtag (dashed lines indicate a breakout)

The Breakouts tool (used with the following parameters: min.size=5, method='multi', beta=.001, degree=1, percent=0.25) detected 14 breakouts (Figure 1), out of which it identified three moments of high user engagement (Table 1).

Table 1. Moments of high user engagement

from	to	tweets	rt	@replies	AVG TPM
07/01 18:07	07/01 23:44	9,194	7,392	150	50.00
08/01 11:42	08/01 23:37	16,048	11,688	472	23.56
09/01 11:55	10/01 00:44	10,159	6,899	465	13.57

Finally, on each subset of tweets created during one of the three moments, we calculated, using the same procedure applied to the entire dataset, a document term matrix of the most frequently used terms. We then grouped those terms according to their co-occurrences.

Table 2. Moments of high user engagement

from	terms	Max sparsity	Most frequent terms
07/01 18:07	5,009	96%	29
08/01 11:42	11,327	96%	22
09/01 11:55	9,735	95%	27

4.DISCUSSION OF ANALYTIC FINDINGS

Adopting the methods suggested in Bruns and Stieglitz's study [4], we used two standard Twitter metrics (i.e. ratio between retweets and tweets and ratio between tweets with URLs over all tweets) to compare #JeNeSuisPasCharlie with other previously studied hashtags. As also discussed in the Literature Review section, Bruns and Stieglitz observed the emergence of two clearly distinct clusters: media events (e.g. #royalwedding, #eurovision) and crisis/emergency events (e.g. #tsunami, #qldflood, #londondriots). In the former case, original tweets are common and URLs are mainly used to share further stories about the media events at hand. In the latter case, during an urgent situation, it is more important to share vital information such as emergency numbers; hence, a characteristically high proportion of retweets and URLs were observed. When mapped on the same chart, the case of #JeNeSuisPasCharlie is noticeably closer to the second cluster characterised by more retweets and more inclusions of URLs (Figure 2).

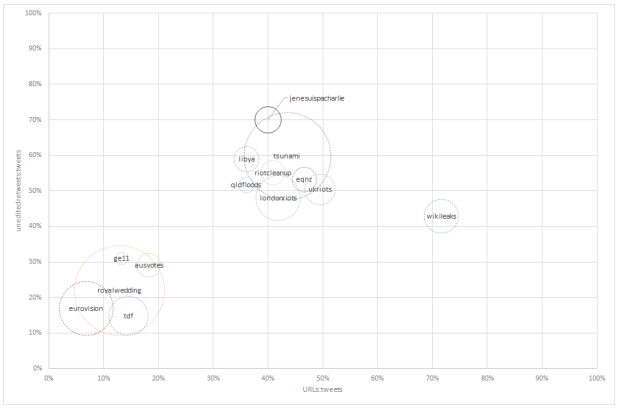


Figure 2. User's activity patterns comparing different Twitter hashtags (size indicates total number of contributor)

A closer analysis of retweets (Table 3) and URLs provided more insights into the nature of #JeNeSuisPasCharlie hashtag.

Table 3. Top 5 most retweeted posts

User	Text of the tweet	N
khurramabad0	Les dessins du dessinateur brésilien Carlos Latuff #JeNeSuisPasCharlie #Charlie_Hebdo #islamophobie http://t.co/a6qrL6pdPt	1,785
RanaHarbi	Last August, The Sydney Morning Herald was forced to remove, apologize for this #JeNeSuisPasCharlie #JeSuisAhmed http://t.co/O7zASRLpD1	868
CoraaantinM	Pr moi ce n'est pas Charlie Hebdo qui est mort mais 2 policiers et des journalistes. L'hommage est à eux, pas au journal #JeNeSuisPasCharlie	794
MaxBlumenthal	A cartoonist with integrity & intellectual consistency – Joe Sacco on Charlie Hebdo #JeNeSuisPasCharlie http://t.co/5uIRwE2wIu	774
SinanLeTurc	Bizarrement quand je dis #JeNeSuisPasCharlie on m'insulte mais quand Charlie insulte notre prophète ça devient de la liberté d'expression.	729

In the aftermath of the shooting, many well-known cartoonists expressed their condolences and solidarity for *Charlie Hebdo* by displaying tribute drawings [20]. Two of the most frequently shared tweets in our dataset also contained links to drawings, but in this case one by the Arab Brazilian freelance political cartoonist Carlos Latuff and another by the Maltese–American cartoonist and journalist Joe Sacco. The two drawings represented a take on the incident that was different from the one put forward by the mainstream community of cartoonists in response to the tragedy of their colleagues at *Charlie Hebdo*. Both Latuff and Sacco pointed out that the magazine had been publishing, in the name of the freedom of speech, images often considered to be offensive for the Muslim population and that the same concept of freedom of speech had not been invoked in the case of an anti-Semitic satire earlier.

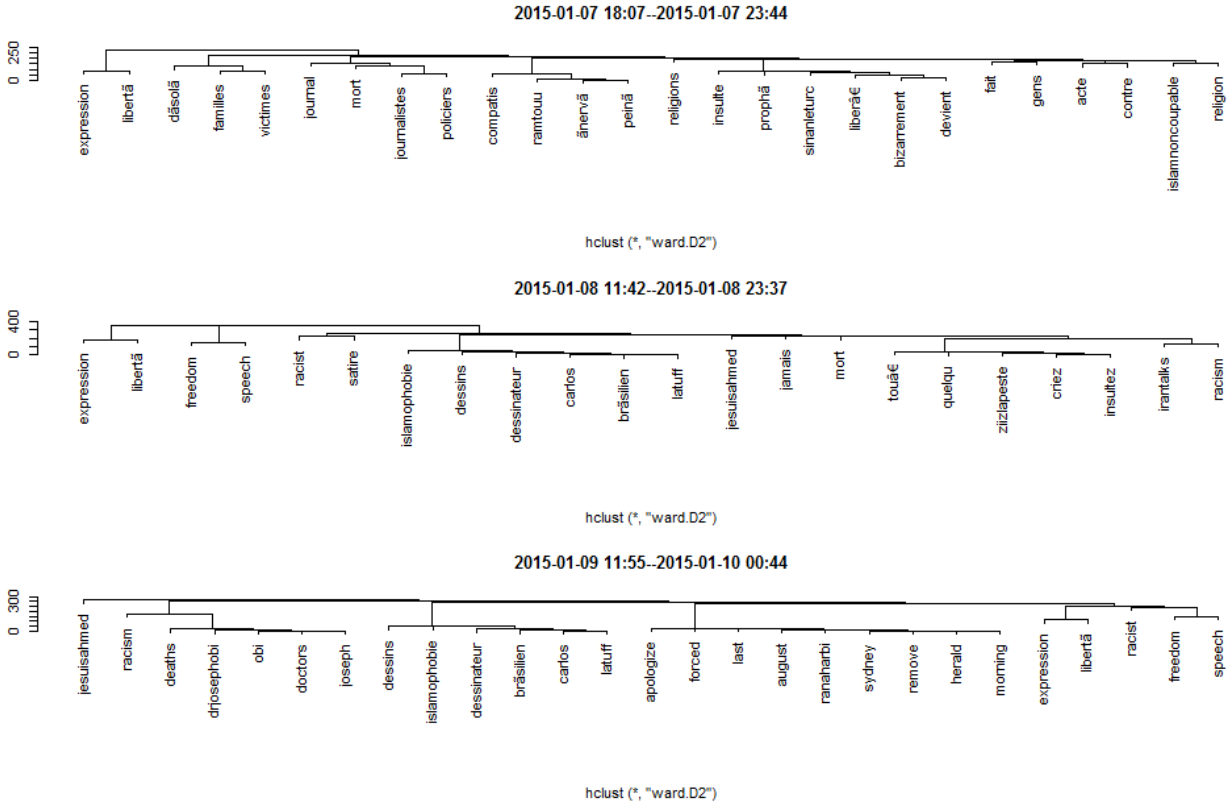


Figure 3. Most frequently used words and their association across the three main phases

Along the same line, another heavily retweeted message recalled the story of Australian newspaper *The Sydney Morning Herald* [15] being forced to issue an apology and remove a drawing that was considered anti-Semitic. This tweet also included the hashtag #JeSuisAhmed, with reference to a Muslim police officer, Ahmed Merabet, also killed during the *Charlie Hebdo* attack. Many Twitter users indeed joined the #JeSuisAhmed hashtag. According to Topsy, it was used over 150,000 times in the days following the attack in a show of condolences for *all* victims of the shooting.

The most frequently shared external sources (URLs) were all images. Links pointing to news sites were rare. This is because #JeNeSuisPasCharlie was not about the news. It's primarily goal was instead to mark and declare an identity by distinction. To that end, 2% of the retrieved tweets were made up of nothing but the hashtag.

As mentioned in the previous section, the first tweet with #JeNeSuisPasCharlie was published less than an hour after what was reported as the first tweet containing #JeSuisCharlie. While the hashtag started as an immediate reaction to #JeSuisCharlie, nevertheless, its nature changed over time.

The Breakout Detection tool developed by Twitter engineers helped us identify three moments of higher user engagement (Table 2). Besides the words related to the most retweeted posts (such as Latuff's cartoon and the *Sydney Morning Herald* case) discussed above, there are a few noteworthy dynamics in Figure 3. First, the clusters of words including *désolé* [sorry] (N=388), *familles* (N=564), *victims* (N=628), and *compatis* [sympathise] (N=409) were present in the first dendrogram but not in the following two. *Liberté* and *expression* (and their corresponding English words) were prominent in all three moments, confirming that the freedom of expression and its contested limits were the real leitmotif across the entire dataset. Terms such as *racism* and *racist* stood out in the second and third moments since users of #JeNeSuisPasCharlie started to approach *Charlie Hebdo*'s satires from different angles than free speech.

5.CONCLUSION

Using a combination of various quantitative techniques, the present study explored the structure of the discussion around the #JeNeSuisPasCharlie hashtag. First, the discussion had a high proportion of retweets (70%) and URLs (41%). Compared to some previously studied hashtags, #JeNeSuisPasCharlie behaved more like crisis/emergency hashtags than media spectacle hashtags. That said, our analytic results also highlighted the heterogeneity of the viewpoints and arguments aggregated under the hashtag in question. Users of the said hashtag showed resistance to the mainstream framing of the *Charlie Hebdo* shooting as the universal value of freedom of expression being threatened by religious intolerance and violence. In this context, retweeting something that would justify their resistance was a way of marking their identity as distinct from what was accepted in the mainstream. Given the sensitivity of the subject, such retweets also helped the users protect themselves from the risk of being viewed as endorsing the violence. We also observed a unique practice of tweeting nothing but the hashtag, amounting to 2% of the dataset. This is a strategy that can be explained in a similar vein.

Over time, there were three distinguished phases in the manifestation of this resistance: Grief (i.e. joining the mourning

for the victims of the attack but indicating a reservation against the proposed frame); Resistance (i.e. starting to voice out the resistance); and Alternatives (i.e. fully developing and deploying alternative frames). In this study, the hashtag was not a conversation marker as previous studies identified but a discursive device that facilitated users to form, enhance, and strategically declare their self-identity.

Our quantitatively oriented methodology here allowed us to identify the topical and network structure of the discussion around #JeNeSuisPasCharlie and its evolution over time. We also suggest as an avenue for further research to delve more qualitatively into the ways in which individual users coped with the sensitive nature of the issue at hand and challenged the mainstream perspective.

6. REFERENCES

[1] Anstead, N. and O'Loughlin, B. 2011. The Emerging Viewertariat and BBC Question Time: Television Debate and Real-Time Commenting Online. *The International Journal of Press/Politics*. 16, 4 (Jul. 2011), 440–462.

[2] Bennett, W.L. and Segerberg, A. 2012. THE LOGIC OF CONNECTIVE ACTION. *Information, Communication and Society*. 15, 5 (2012), 739–768.

[3] Bruns, A. 2012. HOW LONG IS A TWEET? MAPPING DYNAMIC CONVERSATION NETWORKS ON TWITTER USING GAWK AND GEPHI. *Information, Communication and Society*. 15, 9 (2012), 1323–1351.

[4] Bruns, A. and Stieglitz, S. 2013. Towards more systematic Twitter analysis: metrics for tweeting activities. *International journal of social research methodology*. 16, 2 (Mar. 2013), 91–108.

[5] Dubois, E. and Gaffney, D. 2014. The Multiple Facets of Influence: Identifying Political Influentials and Opinion Leaders on Twitter. *The American behavioral scientist*. 58, 10 (Sep. 2014), 1260–1277.

[6] Feinerer, I. et al. 2013. The textcat Package for n-Gram Based Text Categorization in R. *Journal of statistical software*. 52, 6 (Feb. 2013), 1–17.

[7] Giglietto, F. and Selva, D. 2014. Second Screen and Participation: A Content Analysis on a Full Season Dataset of Tweets. *The Journal of communication*. 65, 2 (2014).

[8] Gunnarsson Lorentzen, D. 2014. Polarisation in political Twitter conversations. *Aslib Journal of Information Management*. 66, 3 (2014), 329–341.

[9] Hashtag Activism Isn't a Cop-Out: 2015. *http://www.theatlantic.com/politics/archive/2015/01/not-just-hashtag-activism-why-social-media-matters-to-protestors/384215/*. Accessed: 2015-02-07.

[10] Himelboim, I. and Han, J.Y. 2014. Cancer talk on twitter: community structure and information sources in breast and prostate cancer social networks. *Journal of health communication*. 19, 2 (2014), 210–225.

[11] Hsu, C.-L. and Park, H.W. 2010. Sociology of Hyperlink Networks of Web 1.0, Web 2.0, and Twitter: A Case Study of South Korea. *Social science computer review*. (Sep. 2010).

[12] Israel, Gaza, War & Data: 2014. *https://medium.com/i-data/israel-gaza-war-data-a54969aeb23e*. Accessed: 2015-02-07.

[13] James, N.A. et al. 2014. BreakoutDetection: Breakout Detection via Robust E-Statistics.

[14] Klausen, J. 2015. Tweeting the Jihad: Social Media Networks of Western Foreign Fighters in Syria and Iraq. *Studies in Conflict and Terrorism*. 38, 1 (2015), 1–22.

[15] Meade, A. 2015. SMH cartoon criticised as antisemitic found to breach press council standards. *The Guardian*.

[16] Meyer, D. et al. 2008. Text Mining Infrastructure in R. *Journal of statistical software*. 25, 5 (Mar. 2008), 1–54.

[17] Morstatter, F. et al. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *Seventh International AAAI Conference on Weblogs and Social Media* (Jun. 2013).

[18] Sauter, M. 2014. *The Coming Swarm: DDOS Actions, Hacktivism, and Civil Disobedience on the Internet*. Bloomsbury Academic.

[19] Siapera, E. 2014. Tweeting #Palestine: Twitter and the mediation of Palestine. *International Journal of Cultural Studies*. 17, 6 (Nov. 2014), 539–555.

[20] Telegraph, T.D. 2015. Cartoonists show solidarity after Paris Charlie Hebdo attack, in pictures. *The Daily Telegraph*.

[21] Tufekci, Z. 2014. Online social change: easy to organize, hard to win [Video file]. https://www.ted.com/talks/zeynep_tufekci_how_the_internet_has_made_social_change_easy_to_organize_hard_to_win.. Accessed: 2015-02-07.

[22] Wendling, M. 2015. #JeSuisCharlie creator: Phrase cannot be a trademark. *BBC Trending*.

[23] Xu, W.W. et al. 2014. Predicting Opinion Leaders in Twitter Activism Networks: The Case of the Wisconsin Recall Election. The American behavioural scientist. (Mar. 2014).

A Research Design for the Analysis of the Contemporary Social Movements

Isabel Colucci Coelho, Andrea Lapa
UFSC/Brazil
isabelcolucci,decalapa@gmail.com

Vinicius Ramos
UFRJ/Brazil
vfcr@cos.ufrj.br

Fabio Malini
UFES/Brazil
fabiomalini@gmail.com

ABSTRACT
In the ordinary debate about the political culture decline, social networks have recently changed the social scenario, showing its relevance in down-up social movements. Therefore, social networks are taken here as a potential place for the existence of active citizens - ones that are able and keen about political action in a common world or community. Such recent political revitalization demonstrates the relevance of understanding net activism as a precondition for an active citizenship in the digital culture, where new forms of communication and social interaction seem to influence the democratic relationships in ICT mediated public spheres. The main objective of this article is to present a research design for the identification of elements that promote social empowerment in digital culture. It proposes research procedures for the study of political net activist groups in social networks. Methods, instruments and resources were created and articulated for the collection and treatment of big data and for further qualitative analysis of content, by successive steps of data mining. In addition to contributing to the internet studies field, by proposing a qualitative investigation of social networks, this research design also brings innovation to the Education field as the results of the application of this research design (the identification of important elements for citizens' empowerment) will be used to ground the development of guidelines to teachers and to teachers' education on critical appropriation of social networks in active citizens' education.

Categories and Subject Descriptors
J.4 [**Social and Behavioral Sciences**]: research design for net activism, education and citizen empowerment.

General Terms
Design, Human Factors.

Keywords
quali-quanti methodology; social network analysis; citizen's education; digital culture; net activism.

1.INTRODUCTION
Contemporary societies have witnessed a destabilization of older forms of control and power [1]. Information and Communication Technologies (ICT), especially the political activism catalyzed in social networks, have shown the potential to subvert established power structures and point to alternatives for social transformation.

The latest popular mobilizations that occurred worldwide made vigorous use of social networks in protests and showed a political vitality that calls for a deeper study of this phenomenon (for instance: the Arab Spring - held in many Arab countries, 2010; Occupy Wall Street - USA, 2011; Indignados 15M, 2011 - Spain; June Days - Brazil, 2013; Umbrella Revolution - Hong Kong, 2014).

The political action developed both in social networks and city streets constitute a distinct (and hybrid) public space for democracy, In the debate about the decline of civic and political culture, there are divergent points of view [2]. On one hand, it is argued that the internet trivializes culture and politics, making people not able to carry out meaningful citizen participation [3]. On the other hand, there are optimistic speeches that argue that the internet itself can promote a more inclusive and participatory citizenship (especially among excluded minorities) [4, 5, 6]. Flowing in between, there are varied practices that show the limits of any exact understanding [2].

The political action developed both in social networks and city streets constitute a distinct (and hybrid) public space for democracy, even when recently the general discourse regretted the fading or even the end of politics.

Notwithstanding, to engage young people in politics and civic life again, new means of communication must take place to transcend the limits of traditional politics and also enhance the political dimension of everyday life interests [2]. Arguably, there are new alternatives of political and civic culture under development, which involve more informal methods of participation and collective action that have been disregarded in the attempts to conceptualize political action in actuality [7]. The key question that arises is if this political vitality in social networks could also indicate important elements needed for the empowerment of citizens and their critical education.

Critical education is an educational movement that aims at helping students to develop consciousness of freedom, to recognize authoritarian tendencies, and connect knowledge to power and the ability to take constructive action [8]. The challenge to critical education is described by Hannah Arendt [9], according to whom schools could not promote a critical understanding of the world if they insist in defining a project to the future, which should remain in charge of the new generation. Critical education recognizes the future as a process, a becoming, which depends on these new subjects as authors of their own stories.

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

Therefore, a broad conception of education is required, beyond that which is limited to formal education. Illich [10] is a key reference for reflections on *unschooling*. He argues that traditional schools turned out to stimulate social inequality, especially in poor countries, since it marginalizes those who do not follow it: a class of poor helpless beside an educated elite. More recently, Nóvoa [11] defended a public educational space where many institutions and places take responsibility in education. From this perspective, to overthrow the school walls and recognize the various educative spaces, communication practices in social networks are a legitimate and fertile alternative. Also, these innovative and emancipatory actions take place in defiance of the teachers' own education in the perspective defended here, and still, as adopted in general education system, too instrumental and content-oriented. Nóvoa therefore advocates a revolution in teachers' training to overcome their fragility, which is based on: a) a more open and diverse organization of spaces and times in school; b) a curriculum centered on student learning and not on teaching knowledge and skills; and c) a strongly collaborative pedagogy that uses the networks as communication.

Castells [12] identifies the connection of what he called the autonomy culture of internet with the social movements that emerge in the network, "they partake of a specific culture, the culture of autonomy, the basic cultural matrix of contemporary societies" (p. 167). Thus, the network is a fertile place of research for those who seek an educational model that aims to contribute with emancipation, autonomy and collaboration in the contemporary world. Our intention with such research design is to have the means to analyze data published on Twitter during moments of intense social mobilization, in order to find answers, or at least clues, on how to create more democratic and participatory school practices in digital culture. In other words, how the scenario described above could inspire new critical education models.

2.RESEARCH DESIGN

This work is part of a major project, conducted in Brazil, which investigates how social networks can be used for the education of active citizens. That broad study settles researches in three different contexts: in theory (grounded in Critical Thinking); in case studies of net activism; and in pedagogical practices using social networks (at elementary and high schools, and also at universities).

The research design presented here is restricted to the observation of net activism in social networks, which integrates the case study context along with interviews held with key actors of those net activist groups (not presented here). It was originally created for the study of the Free Pass Movement's political action. The Free Pass Movement (Movimento Passe Livre - MPL in Portuguese) is an activist group that advocates for free public transportation across Brazil and presents itself as an autonomous, nonpartisan, independent and horizontal social movement. Five main factors led to the selection of MPL's political action the object of this research:

- The maturity of the group, that has over ten years of existence;
- Its derivation into horizontally organized groups, active throughout the country by a federative model;
- Their final object of claim: the right to the city, or to the public space;
- Its strong presence in the public space as the virtual networks;
- Its crucial role for the start of the June Days - public demonstrations that dragged millions of people to Brazilian streets in June of 2013.

At the moment, we are currently applying this research design to mine 70,000 posts published on Twitter during the first month of the protests (June to July 2013), with the term "Passe Livre".

The major challenge of this research is the qualitative data analysis, since it deals with large volumes and varieties of data that are produced in a high speed in social networks. The unfeasibility of a manual, laborious and time-consuming process of analysis has been overcome through a partnership with Labic/ UFES and the adaptation to the context and objectives of this research, of the Perspectival Method of Network Analysis (PMNA) developed by them [13]. Although their method makes use of automated and quantitative data treatment, the manageable data resulted of these processes allows a qualitative analysis.

The Perspectival Method of Network Analysis is grounded on the fundamentals of Complex Network Theory and aims to demonstrate the different points of view that rise within a topic of politic mobilization on social networks. PMNA was crucial for the comprehension of the many clusters of ideological positioning existing during the demonstrations that took place in Brazil, in June 2013. The analysis of retweeted messages with the hashtag #vemprarua[1] allowed the identification of seven major points of view: activism, hacking, media, politic mocking, human rights, clicktivism and fandoms. The method brings to light the idea that networks on Twitter are not an entire body, but are side-by-side parts [13].

With PMNA, it is possible to handle posts exchanged in social networks in successive stages of extraction, mining, processing and visualization of large volumes of data. With the data resulted from this method, we were able to add new steps of observation, according to categories derived from previous phases of our own research. As a result, we developed a model of investigation that allows the discursive analysis from posts generated on net activist groups.

Three new steps to analyze the data provided by PMNA were added, in order to obtain more in-depth qualitative results: (a) first, the creation of procedures to identify the moments within the dataset with potential to reveal the process that should be observed according to the research purposes. These specific contexts are named *Spaces of Possibility.* In our case study, they represent moments of dialog and social interaction; (b) secondly, the analysis of the *Spaces of Possibility* identified in the previous stage, in order to find examples of predetermined categories, brought from the review of literature. These categories are the social process we aim at observing and are referred to as *Relevant Processes* in this research design. The two procedures described above substantially reduce the amount of data to be analyzed, and they also make it possible to retrieve the relevant political dialog in the dataset in a viable quantity for qualitative analysis; c) in the final step, we gathered the dialog thread of the selected potential posts and started the content analysis.

The procedures briefly introduced above are detailed as follows: data collection (Section 2.1); data treatment, comprising the mining procedures, processing and visualization (Section 2.2); and data analysis (Section 2.3).

[1] The hashtag #vemprarua, in Portuguese could be translated to "come to the streets".

2.1.Data Collection

The posts are collected by a search and monitor engine that filters the Twitter stream by keywords and hashtags, and stores the data in a CSV format text file. Similar tools are in market today, such as Topsy (http://topsy.com) and Flocker (http://flocker.outliers.es). Most of these engines are built on top of the Twitter API (https://dev.twitter.com/overview/api). For example, to capture tweets based on the hashtag "#brasil", the software captures all tweets containing the terms "#brasil" at the same time it is being posted (not allowing past time data collection) and stores these data into a dataset that contains: a) UserID, the identification of the user that sent the tweet; b) Time, the date and time that the tweet was posted; c) Tweet Text, the tweet's content; d) Geolocation, the geographic location of the user (only if the user agreed to share it); e) Image, if it is tweeted an image its location is stored as an URL.

After this phase of extraction, a script developed as part of PMNA is run and the dataset is processed to create 20 different text files, in which each of them contains different statistics about the tweets. The script, written in python language, is open and free for usage and modification according to one's own purpose[2].

These files are organized according to the post date, hashtags used, user activity, locations and other criteria. In our research, two of them have substantial importance: "top words" (the relation of the one thousand words more frequently used in posts containing the selected hashtag or keyword); and the "top hashtags" file (set of one thousand hashtags most commonly associated with the hashtags and words used in the dataset).

Three other files from the dataset were also created, containing a sample of one third of the full amount of tweets, collected from the beginning, the middle and the end of the posts collection. Thus, our research deals with data organized in five text files: the three sample selection, plus the "top words" and "top hashtag" files.

2.2.Data Treatment

Due to the large quantity of data and the necessity of a qualitative filter to reduce it to a suitable amount, the data treatment is separated in three steps: (a) Data Mining through *Spaces of Possibility*; (b) Data Mining through *Relevant Processes*; (c) and Compilation of Dialogs.

2.2.1.Data Mining through Spaces of Possibility

The first methodological procedure (added to PMNA) is to identify the categories in the dataset which can possibly contain relevant processes for observation. These categories were called *Spaces of Possibility*, since they hold the potential of occurrence of the processes that are relevant to the study. For our research purposes, our theoretical grounding suggests that these would be moments of dialog, social integration, conflict and debate among controversial issues, as well as confluence of online and offline action.

First Step: Identification of categories

OBJECTIVE: To define terms and words for a first filtering of the dataset. To determinate what words and terms indicate the existence of a *space of possibility*.

DATA PROCESSING: a) manual reading of the five documents (three samples of tweets, topwords and tophashtags) to select words and terms that are often used in posts that can potentially show the social processes the present research aims at investigating; b) discussion and alignment among researchers about the election of categories and their consensus about meaning.

PRODUCT: A library of terms and words linked to the *spaces of possibility*. For instance, to identify dialogs, we select terms that demonstrate exchange and sharing of ideas, personal exposure, absence of hierarchy or leadership, multiple authorship, opposition of ideas, conflict, use of the first person (singular), which can present willingness to negotiate different points of view.

Table 1 - Example of the Library of Terms and Words linked to *spaces of possibility* in the MPL case study

Dialog	Social Integration	Online/Offline confluence
Conversation; Open to talk; Did you know that...; Meeting; Report; Against/ For; Let's talk; Comment; Debate/ Idea.	Represents me; Group; Fighting; Supporting; Fear; Collective/ Friends; Volunteers; Help; Support; We are/ We all.	Live now; Protest; Occupy; Happening now; On the street/In front of; Itinerary/Walk; Meeting point; Now; It's today;

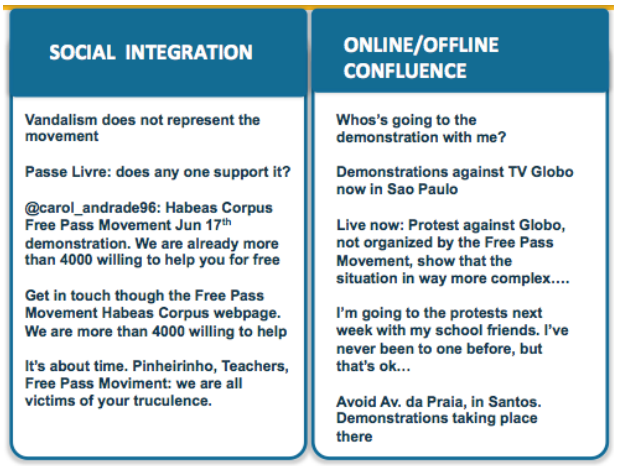
Second Step: Mining for Spaces of Possibility

OBJECTIVE: To select and separate posts that trigger processes, spaces where there is a probability of finding political action and, with it, elements that promoted it. From the library of terms and words that identify *spaces of possibility*, a script is run to filter the entire dataset through the categories of *spaces of possibility* defined in the previous phase.

DATA PROCESSING: a) adaptation, testing and application of PMNA data mining script; b) filtering the whole dataset by categories (defined in the library of terms and words).

PRODUCT: Graphical interface with featured posts nestled by the selected categories.

Picture 1 Graphical interface - posts nestled by spaces of possibility (posts literally translated from Portuguese)



[2] It can be found at https://github.com/ufeslabic/parse-tweets.

In the picture above, we see posts that demonstrate social integration and Online/Offline confluence. In social integration, for instance, we bring posts regarding the union of a community of over 4,000 lawyers volunteering to obtain *habeas corpus* (great writ) for the detained protester. The other posts demonstrate a sense of belonging to a social group. In Online/Offline Confluence, we have examples of posts that refer to events taking place on the streets.

2.2.2.Data Mining: Relevant Processes

First Step: Identification of categories

OBJECTIVE: To select relevant social processes that may harbor the political action to be investigated. Bring from reviews of literature some predetermined analytical categories as *relevant processes* that may guide the identification of posts to be studied.

DATA PROCESSING: a) To develop indicators and metrics to identify these processes in the *spaces of possibility*. Three *relevant processes*, relating to our object of research, were highlighted to demonstrate the proposed data treatment in this phase, as shown in Table 2:

Second Step: Mining through Processes

OBJECTIVE: To identify where/whether the searched processes existed in those *spaces of possibility*. To select, highlight and separate them. The purpose of this step is to identify some potential posts within the *spaces of possibility* and extract the dialog it may have generated for analysis.

DATA PROCESSING: a) Manual analysis of the dataset; b) To mine the posts with a script according to the metric that was defined in the previous step and separate them into the pre-selected categories.
PRODUCT: Graphical interface with featured posts nestled by *Relevant Processes* (similar to the interface presented for *spaces of possibility*).

2.3.COMPILATION OF DIALOGS

OBJECTIVE: From the potential posts identified in the previous phase, the *relevant processes,* find and bring the thread of this post. That means, from the selected post (that is yet a fragment), bring the other posts that were generated from it, as mention, retweet or response to it.

DATA PROCESSING: a) to develop and run a script to collect other posts connected to the one selected. b) To bring the threads of the dialogue that have come to light from potential posts selected in *relevant processes*.

PRODUCT: Document containing the threads of all messages regarding the selected posts, separated by analytical category (*Relevant Processes*).

2.4.DATA ANALYSIS

At this stage, with an adequate amount of data, it is proposed a more qualitative approach with an in-depth, interpretative and inferential analysis. From the selection of the dialogs, the objective is to identify what, in the *Space of Possibility* and the occurrence of *Relevant Processes*, allowed and promoted the existence of an active citizen. This phase deals with an immersion in the data to pick up clues for critical education. In other words, it makes meaning and understanding out of the *relevant processes* in terms of participants' definition of situations, important themes, elements that may have generated or promoted the political action in social networks.

Table 2 - Table of Analytic Categories developed by Andrea Lapa, Isabel Coelho, Simone Schwertl, Andreson Lopes.

Relevant Process: Plurality	
Description	**Indicators**
Constitutes the public. Welcomes the individual's singularities on equal terms. It has two aspects: a) Equality - we are all equal; and b) Distinction - the uniqueness of each person revealed by discourse and action [14].	Shared space of exchanging ideas (equality) (visible and audible active beings); diversity of perspectives in the debate (distinction); welcome in the group (and authorities support) of various perspectives that are included in the debate.
Relevant Process: Communicative Action	
Description	**Indicators**
There is no goal to be achieved, but the agreement between participating subjects, that is, all those involved in the dialog are considered qualified to interfere in the process. The language is not used as a mean of transmission of information (strategic action) but as a source of social integration (communicative action) [15].	Motivation for understanding; language used as a source of social integration (search for dialog, exchange - to generate the debate that leads to an agreement); argumentative exchange between the published messages; search for a common sense, not just the exposure of individual understandings.
Relevant Process: Common World	
Description	**Indicators**
World of shared existence. Participant actors try cooperatively to define their action plans, taking into account each other, the horizon of a shared world in the basis of a common interpretation of situations [15]. Most people enter a social movement with their own goals and motivations, and come to find common denominators in the movement's practice itself.	Sense of inclusion in the group; move from an individual vision to a collective one.

The data analysis is a content analysis of the dialogs. There are several computer packages for qualitative data [16], for example: AQUAD; ATLAS.ti; Nvivo; Textbase Alpha, ETHNOGRAPH, which do not perform the analysis but can assist it by organizing and structuring text for subsequent analysis. For the MPL case study we chose WebQDA, which allows features such as to search for text, codes, nodes and categories; to organize and filter them presenting grouped data according to criterion desired; to run boolean and proximity searches; to present data in sequences and locate the text in surrounding material providing context; to classify subjects and subsets, to enable memos and also treatment of non numerical and unstructured data; and to question data by

crossing categories, codes and subjects. Most of all, due to the collaborative work of analysis through cloud computing that counted on the presence of many Brazilian researchers, the software privileged runs in Portuguese, the same language as in the database.

This analysis phase involves: coding, categorizing (creating meaningful categories into which the units of analysis – tweet posts – can be placed), comparing (categories and making links between them), and concluding (drawing theoretical conclusions from the text). The codes (*Relevant Processes* framework), organize the understanding of the problem in a specific context that structures the categorization, which is exploratory and based on emergent themes and patterns. The description of the phenomena, its association with other categories, and the identification of relations between variables allow the development of interpretation from the social interactions in the dataset.

The final product of this research design is a provisional guide of important elements for the critical education of active citizens in social networks. Such guide provides effective recommendations that should be pertinent, if not fundamental, to: teachers' education, pedagogical application by teachers and educators, action research in teaching practices (the other contexts of the research project - formal and non-formal education), and to orient the other stage of the case study (interviews).

3.CONCLUSION

This work is part of the investigative efforts of researchers in the Education field who seek alternatives for critical education in digital culture. In the debate about the decline of political culture, political action on social networks presents alternatives to the traditional education system. In this perspective, the interactions that take place in social networks are, perhaps, a precondition for citizenship in digital culture, where democratic relations are promoted in public spheres and by new forms of online participation.

The acquaintance of elements that can promote the existence and the empowerment of citizens is a demand for an emancipatory education of the XXI century. Although education plays an important role in this scenario, teachers and educators lack references and abilities to empower active citizens and enhance a critical education in digital culture. The research presented in this article is a contribution in this direction. It deals with the development of a research design for the investigation of net activist groups aiming to identify elements that promote political action in social networks.

Methods and instruments for data collection, data treatment, and data analysis were articulated and created. For the extraction and data treatment, it was presented a model, which was adapted to the context of this research, the Perspectival Method of Network Analysis [13]. In the steps of mining the large amount of data, there was the definition of the categories of analysis in two steps: *Spaces of Possibility* and *Relevant Processes*. At the end, it is possible to extract the discursive exchange for the final stage of content analysis of the dialogs (*Compilation of Dialogs*).

The research objective is to guide teachers and educators in their practices inside and outside of school. The final result of the investigation conducted with this instrument will provide some guidelines for teachers' education, counting on the reference of relevant elements for critical education in digital culture. In a manual analysis of the dataset we could foresee some provision of results. For instance, after identifying plurality (the existence of diverse ideas) in a space of possibility (social integration), it showed an important mediation of some key actors, not a single one. In the content analysis of the dialogue, their role can show to teachers how to promote these elements in educational practice.

4.ACKNOWLEDGMENTS

This work was developed as part of the project "Education and Technology: investigating the potential of social network for the active citizen's education" (Comunic/UFSC/Brazil). It is funded by CAPES, through the postdoctoral fellowship of Andrea Lapa and also the support of the project "Public Politics and Education Networks (RPPE)" which is coordinated by Tamara Egler. We are grateful for the partnership with Labic (Laboratório de Estudos sobre Imagem e Cibercultura /UFES) and also the research design review of Stefano Renzi, Jane Klobas and Michel Menou.

5.REFERENCES

1. FORTUNATI, L. (2014). Media Between Power and Empowerment; Can we resolve the dilemma? The Information Society, 30: 169-183, 2014.
2. BUCKINGHAM, D.; BADAJI, S. (2013). The civic web: young people, the Internet and civic participation. The John D. and Catherine T. MacArthur Foundation series on digital media and learning. MIT Press, Cambridge, MA, USA.
3. KEEN, A. (2007) The cult of the amateur: how today's internet is killing our culture. Nova York: Doubleday/ Currency
4. BENKLER, Y. (2006) The wealth of networks: how social production transforms markets and freedom. Yale University Press - New Haven and London.
5. RHEINGOLD, H. (2007) Smart mobs: The next social revolution. Basic books.
6. SHIRKY, C. (2009). Here Comes Everybody: How Change Happens when People Come Together. New York: Penguin.
7. DAHLGREN, P. (2004). Civic cultures and net activism: modest hopes for the EU public sphere. Conference on One EU - Many Publics? At University of Stirling, 5-6 Feb 2004
8. GIROUX, H. (2010). Lessons From Paulo Freire. Chronicle of Higher Education. 10/22/2010, Vol. 57 Issue 9, pB15-B16. 2p
9. ARENDT, H. (1977). ''The Crisis in Education'' in between past and future. New York: Penguin.
10. ILLICH, I. (1971) Deschooling Society. New York: Harper and Row.
11. NÓVOA, A (2014). Nada será como antes. Revista Pátio-Ensino Fundamental: "O futuro da sala de aula" (Entrevista) no. 72. Porto Alegre: Artmed Press, Novembro 2014
12. CASTELLS, M (2013). Redes de Indignação e Esperança (Networks of Outrage and Hope: Social Movements in the Internet Age). Rio de Janeiro: Zahar.
13. MALINI, F.; CALMON, P.; MEDEIROS, J.; MALINI, M. (2015) "Multiple points of view in #VemPraRua ReTweets: the perspectival method of network analysis". Conference Twitter for Research, Lyon, 22 - 23 April 2015.
14. ARENDT, H. (1998). The Human Condition. The University of Chicago Press, Chicago, IL, USA
15. HABERMAS, J. (1994) Postmetaphysical Thinking: Philosophical Essays. MIT Press
16. KELLE, U. (ed.) (1995) Computer-Aided Qualitative Data Analysis. London: Sage.

Section IV:

Named Entity Extraction & Linking (NEEL) Challenge

edited by
Amparo Elizabeth Cano Basave, Giuseppe Rizzo, Andrea Varga & Bianca Pereira
Matthew Rowe, Milan Stankovic & Aba-Sah Dadzie

Making Sense of Microposts (#Microposts2015) Named Entity rEcognition & Linking Challenge

Giuseppe Rizzo
EURECOM, France
giuseppe.rizzo@eurecom.fr

Amparo E. Cano
KMi, The Open University, UK
amparo.cano@open.ac.uk

Bianca Pereira
Insight Centre for Data Analytics, Ireland
bianca.pereira@insight-centre.org

Andrea Varga
Swiss Re, London, UK
varga.andy@gmail.com

ABSTRACT

Microposts are small fragments of social media content and a popular medium for sharing facts, opinions and emotions. Collectively, they comprise a wealth of data that is increasing exponentially, and which therefore presents new challenges for the Information Extraction community, among others. This paper describes the *Making Sense of Microposts* (#Microposts2015) Workshop's **Named Entity rEcognition and Linking (NEEL)** Challenge, held as part of the 2015 World Wide Web conference (WWW'15). The challenge task comprised automatic recognition and linking of entities appearing in different event streams of English Microposts on Twitter. Participants were set the task of investigating novel strategies for extracting entities in a tweet stream, typing these based on a set of pre-defined classes, and linking to DBpedia or NIL referents. They were also asked to implement a web service to run their systems, to minimize human involvement in the evaluation and allow measuring of processing times. The challenge attracted a lot of interest: 29 research groups expressed an intent to participate, out of which 21 signed the agreement required to be given a copy of the training and development datasets. Seven teams participated in the final evaluation of the challenge task, out of which six completed all requirements, including submission of an abstract describing their approach. The submissions covered sequential and joint linguistic methods, end-to-end and hybrid end-to-end, and linguistic approaches for tackling the challenge task. We describe the evaluation process and discuss the performance of the different approaches to the #Microposts2015 NEEL Challenge. We also release, with this paper, the #Microposts2015 NEEL Challenge Gold Standard, comprising the set of manually annotated tweets.

Keywords

Microposts, Named Entity Recognition, Named Entity Linking, Disambiguation, Knowledge Base, Evaluation, Challenge

1. INTRODUCTION

Microposts are short text messages published using minimal effort via social media platforms. They provide a publicly accessible wealth of data which has proven to be useful in different applications and contexts (e.g., music recommendation, social bots, spam detection, emergency response). However, extracting data from Microposts and linking it to external sources presents various challenges, due, among others, to the inherent characteristics of this type of data:

i) the restricted length;
ii) the noisy lexical nature, where terminology differs between users when referring to the same thing, and non-standard abbreviations are common.

A commonly used approach for making sense of Microposts is the use of textual cues, which provide contextual features for the underlying tweet content. One example of such a cue is the use of *Named Entities*. Extracting named entities from Microposts has, however, proven to be a challenging task; this was the focus of the Concept Extraction (CE) Challenge, part of the 2013 workshop, #MSM2013 [4]. A step further into the use of such cues is to ground entities in tweets by linking them to Knowledge Base referents. This prompted the Named Entity Extraction and Linking (NEEL) Challenge the following year, in #Microposts2014 [3]. These two research avenues, which add to the intrinsic complexity of the tasks proposed in 2013 and '14, prompted the Named Entity rEcognition and Linking (NEEL) Challenge in #Microposts2015. In NEEL 2015 we investigated further the role of the named entity type in the process, and the identification of named entities that cannot be grounded because they do not have a Knowledge Base referent. The English DBpedia 2014[1] dataset was the designated reference Knowledge Base for the 2015 NEEL challenge.

From the first Concept Extraction challenge (in 2013) through to the 2015 NEEL challenge, we have received over 40 submissions proposing state of the art approaches for extracting, typing, linking, and clustering relevant pieces of data from Microposts, namely, named entities. The purpose of each challenge was to set up an open and competitive environment that would encourage participants to deliver novel or improve on existing approaches for recognizing and linking entities from Microposts to either a reference Knowledge Base entry or NIL where such a reference does not exist. To encourage competition we solicited sponsorship for the winning submission, an award of €1,500. This was provided by SpazioDati,[2] a startup operating in the Big Data & Semantic Web market, who are active in the research community of entity linking.

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

[1] http://wiki.dbpedia.org
[2] http://www.spaziodati.eu

This generous sponsorship is testament to the growing interest in challenges related to automatic approaches for gleaning information from (the very large amounts of) social media data generated across all aspects of life, and whose knowledge content is recognised to be of value to industry.

This paper describes the #Microposts2015 NEEL Challenge, detailing its rationale and research challenges, the collaborative annotation of the corpus of Microposts, and our evaluation of the performance of each submission. We describe the approaches taken in the participants' systems – which use both established and novel, alternative approaches to entity extraction, typing, linking and clustering. The resulting body of work has implications for researchers, application designers and social media engineers who wish to harvest information from Microposts for their own objectives.

2. TASK DEFINITION AND EVALUATION

In this section we describe the goal of the challenge, the task set, and the process we followed to generate the corpus of Microposts.

2.1 The Task and Research Challenges

The 2015 challenge required participants to build automated systems to solve three main tasks:

i) extraction and typing of entity mentions within a tweet;
ii) linking of each mention to a referent in the English DBpedia 2014 dataset representing the same real world entity, or NIL for cases where no such entry exists;
iii) clustering of each unique, non-linked entity to a NIL identifier, where each cluster contains only mentions to the same real world entity.

In the rest of this paper we refer to the term appearing in a text as either an *entity mention* or simply an *entity*, while we refer to its DBpedia referent as the *candidate*. Consequently, the operation of entity detection is also referred to as *mention detection*, whilst for entity linking we use *candidate selection*.

An entity, in the context of this challenge, is used in the general sense of being, not requiring a material existence but only to be an instance of a taxonomy class. Thus, a mention of an entity in a tweet can be seen as a proper noun or an acronym. The extent of an entity is the entire string representing the name, excluding the preceding definite article (i.e., "the") and any other pre-posed (e.g., "Dr.", "Mr.") or post-posed modifiers.

In this task we consider an entity to be referenced in a tweet as a proper noun or an acronym when: i) it belongs to one of the categories specified in the NEEL Taxonomy (see Appendix A); and ii) it can be linked to an English DBpedia referent or to a NIL reference given the context of the tweet.

Pronouns (e.g., he/she, him/her) are not considered mentions of entities in the context of this challenge. Lowercase and compressed words (e.g., "c u 2night" rather than "see you tonight") are common in tweets. Thus, they are still considered mentions if they can be directly mapped to proper nouns. Complete entity extents, and not their substrings, are considered a valid mention. For example, from the following text excerpt: "Barack Obama gives a speech at NATO", neither of the words *Barack* nor *Obama* is considered by themselves, but rather *Barack Obama*. This is because they constitute a substring of the full mention [`Barack Obama`]. However, in the text: "Barack was born in the city, at which time his parents named him Obama" each of the terms [`Barack`] and [`Obama`] should be selected as a separate entity mention.

Nested entities with qualifiers should be considered as independent entities; similarly, compound entities should be annotated in isolation. E.g.,

Tweet:
`Alabama CF Taylor Dugas has decided to end negotiations with the Cubs and will return to Alabama for his senior season. #bamabaseball`

For this tweet, the [`Alabama CF`] entity qualifies [`Taylor Dugas`]; the annotation for such a case should be: [`Alabama CF, Organization, dbp:Alabama_Crimson_Tide`] and [`Taylor Dugas, Person, NIL1`], where NIL1[3] is the unique NIL identifier describing the real world entity "Taylor Dugas".

2.1.1 *Noun phrases completing the definition of an entity*

In the 2015 challenge, as opposed to the previous edition, not all noun phrases are considered as entity mentions. E.g., in:

Tweet:
`I am happy that an #asian team have won the womens world cup! After just returning from #asia i have seen how special you all are! Congrats`

While "asian team" could be considered as an Organization-type it can refer to multiple entities. Therefore we do not consider it as an entity mention, and it should not be annotated.

While noun phrases can be linked to existing entities, we do not consider them as entity mentions. In such cases we only keep "embedded" entity mentions. E.g., in:

Tweet:
`head of sharm el sheikh hospital is DENYING`

"head of sharm el sheikh hospital" refers to a Person-type; however, since it is not a proper noun we do not consider it as an entity mention. For that reason, in this case the annotation should only contain the embedded entity [`sharm el sheikh hospital`]: [`sharm el sheikh hospital, Organization, dbp:Sharm_International_Hospital`].

In the tweet:

[3]NIL1 is composed of two parts: NIL and the suffix 1. Any suffix, numeric or alphanumeric, is considered as a valid suffix.

Tweet:
`The best Panasonic LUMIX digital camera from a wide range of models`

while digital camera describes the entity "Panasonic LUMIX", it is not considered within the entity annotation, since it is used in the context as a noun phrase.[4] In this case the annotation should be [`Panasonic, ORG, dbp:Panasonic`] [`LUMIX, Product, dbp:Lumix`].

Entity mentions in a tweet can also be typified based on the context in which they are used. In:

Tweet:
`Five New Apple Retail Stores Opening Around the World: As we reported, Apple is opening 5 new retail stores on ...`

In this case [`Apple Retail Stores`] refers to a Location-type, while the second [`Apple`] mention refers to an Organisation-type.

2.1.2 Special Cases in Social Media (# and @)

Entities may be referenced in a tweet preceded or composed by # and @, e.g.:

Tweets:
`#[Obama] is proud to support the Respect for Marriage Act.`
`#[Barack Obama] is proud to support the Respect for Marriage Act.`
`@[BarackObama] is proud to support the Respect for Marriage Act.`

Hashtags (i.e., words referenced by a #) can refer to entities, but this does not mean that all hashtags will be considered as entities. Further, for our purposes, the characters # and @ should not be included in the annotation string. We consider the following cases:

Hashtagged nouns and noun-phrases:

Tweet:
`I burned the cake again. #fail`

The hashtag "#fail" does not represent an entity. Thus, it should not be annotated as an entity mention.

Partially tagged entities:

[4]Panasonic LUMIX refers to a series of cameras. Therefore to be considered a proper noun it should be followed by a number or an identifier.

Tweet:
`Congrats to Wayne Gretzky, his son Trevor has officially signed with the Chicago @Cubs today`

Here "Chicago @Cubs" refers to the proper noun characterising the [`Chicago Cubs`] entity. (Note that in this case "Chicago" is not a qualifier, but rather, part of the entity mention.) The annotation should therefore be [`Chicago, Organization, dbp:Chicago_Cubs`] and [`Cubs, Organization, dbp:Chicago_Cubs`].

Tagged entities:

If a proper noun is split and tagged with two hashtags, the entity mention should be split into two separate mentions.

Tweet:
`#Amy #Winehouse`

In this case we annotate [`Amy, Person, dbp:Amy_Winehouse`] [`Winehouse, Person, dbp:Amy_Winehouse`]

2.1.3 Use of Nicknames

The use of nicknames (i.e., descriptive names replacing the actual name of an entity) are commonplace in Social Media, e.g., the use of "SFGiants" to refer to "the San Francisco Giants". For these cases, nicknames are co-referenced to the entity they refer to in the context of a tweet.

Tweet:
`#[Panda] with 3 straight hits to give #[SFGiants] 6-1 lead in 12th`

We annotate [`Panda, Person, dbp:Pablo_Sandoval`] and [`SFGiants, Organization, dbp:San_Francisco_Giants`].

2.2 Evaluation Strategy

Participants were required to implement their systems as a publicly accessible web service following a REST-based protocol, in order to submit (up to 10) contending entries to a registry of the NEEL challenge services. In this context, we refer to a contending entry as the participant's REST endpoint queried in the evaluation campaign. Each endpoint had a Web address (URI) and a name, which we defined as run_{ID}. Upon receiving the registration of the REST endpoint, calls to the contending entry were scheduled in two different time windows, namely, D-Time – to test the APIs, and T-Time – for the final evaluation and metric computations. To ensure correctness of the results and avoid any loss we triggered a large number of queries and statistically evaluated the results.

2.2.1 Metrics and Scorer

The evaluation was conducted using four different metrics:

i) strong_typed_mention_match,
ii) strong_link_match,
iii) mention_ceaf,
iv) latency.

The *strong_typed_mention_match* evaluates the micro average F_1 score for all annotations considering the mention boundaries and their types. The *strong_link_match* is the micro average F_1 score for annotations considering the correct link for each mention. The *mention_ceaf* (Constrained Entity-Alignment F-measure) [10] is a clustering metric developed to evaluate clusters of annotations. It evaluates the F_1 score for both NIL and non-NIL annotations in a set of mentions. The *latency* measures the computation time of an entry (in seconds), to annotate a tweet. The final score is computed according to Equation 1. The *latency* metric was included only to resolve cases where there was a tie in the evaluation score.

$$\begin{aligned} score = {} & 0.4 * mention\_ceaf \\ & + 0.3 * strong\_typed\_mention\_match \\ & + 0.3 * strong\_link\_match \end{aligned} \quad (1)$$

The scorer proposed for the TAC KBP 2014 task[5] was used to perform the evaluation.

2.2.2 Selection of the Annotation Results

Algorithm 1 EVALUATE($E, Tweet, N = 100, M = 30$)

for all $e_i \in E$ **do**
  $A^S = \emptyset, L^S = \emptyset$
  for all $t_j \in Tweet$ **do**
    for all $n_k \in N$ **do**
      $(A, L) = annotate(t_j, e_i)$
    end for

    // Majority Voting Selection of a from A
    for all $a_k \in A$ **do**
      $hash(a_k)$
    end for
    A_j^S = Majority Voting on the exact same $hash(a_k)$

    // Random Selection of l from L
    generate L^T from the uniformly random selection of M l from L
    (μ, σ) = computeMuAndSigma(L^T)
    $L_j^S = (\mu, \sigma)$
  end for
end for

To ensure the correctness of the results and avoid any loss we triggered N (with N=100) calls to each entry. We then applied a majority voting approach over the set of annotations per tweet and statistically evaluated the latency by applying the law of large numbers [14]. Algorithm 1 provides a sketch of the algorithm used during the evaluation campaign.

3. PARTICIPANT OVERVIEW

The challenge attracted a lot of interest from research groups spread around the world. Twenty-nine groups expressed their intent to participate in the challenge; out of which twenty-one signed the agreement required to be given a copy of the training and development datasets. Seven teams participated in the final evaluation of the challenge task, out of which six completed submission with an abstract describing the approach they took. The final submissions are listed in Table 1.

Table 2 provides a taxonomy of the approaches proposed this year for tackling the challenge task. From an historical perspective, starting from the first Concept Extraction (CE) challenge till the current, 2015, apart from the NIL detection and clustering introduced in this challenge, we observed:

1. the consolidation of a normalization procedure, namely preprocessing, to increase the expressiveness of the tweets, e.g. via expansion of Twitter accounts and hashtags with the actual names of entities they represent;
2. the consolidated contribution of Knowledge Bases in the Mention Detection and Typing task. This leads to higher coverage, which, along with the linguistic analysis and type prediction, better fits the Microposts domain;
3. the consolidation of the Candidate Selection performed as an End-to-End approach. Such an approach has been further developed with the addition of fuzzy distance functions operating over n-grams and acronyms;
4. a considerable decrease in off-the-shelf systems.

We provide next a detailed description of each contribution.

In [15], Yamada et al., present a five-sequential stage approach: preprocessing, generation of potential entity mentions, candidate selection, NIL detection, and entity mention typing. In the preprocessing stage, they propose a tokenization and Part-of-Speech (POS) tagging approach based on [7], along with the extraction of tweet timestamps. They tackle the generation of potential entity mentions by computing n-grams (with $n = 1..10$ words) and matching them to Wikipedia titles, Wikipedia titles of the redirect pages, and anchor text using exact, fuzzy, and approximate match functions. An in-house dictionary of acronyms is built by splitting the mention surface into different n-grams (where 1 n-gram corresponds to 1 char). At this stage all entity mentions are linked to their candidates, i.e., the Wikipedia counterparts. The candidate selection is approached as a learning to rank problem: to each mention is assigned a confidence score computed as the output of a supervised learning approach using Random Forest as the classifier. An empirically defined threshold is used to select the relevant mentions; in the case of mention overlap the span with the highest score is selected. The NIL detection is tackled as a supervised learning task, in which Random Forest is used. The features used are the predicted entity types, contextual features such as surrounding words, POS, length of the n-gram and capitalization features. The mention entity typing stage is treated as a supervised learning task where two independent classifiers are built: a Logistic Regression classifier for typing entity mentions and a Random Forest for typing NIL entries.

Gârbacea et al., [6] present a sequential approach composed of four stages: entity mention detection, candidate selection, NIL clustering, and resolution of overlapping mentions. The first stage is tackled by empowering both an annotation-based off-the-shelf system, Semanticizer,[6] and a Named Entity Recognition classifier trained using the challenge dataset. For each entity mention, a Learning to Rank supervised model is used to select the most representative DBpedia reference of the entity mention (candidate detection). The resulting type of the DBPedia reference entity is used to type the

[5] https://github.com/wikilinks/neleval/wiki/Evaluation

[6] https://github.com/semanticize/semanticizer

Table 1: Accepted submissions with team affiliations and number of runs for each.

Reference	Team's affiliation	Team Name	Authors	No. of entries
[15]	Studio Ousia and Keio University and National Institute of Informatics	ousia	Yamada *et al.*	10
[6]	University of Amsterdam	uva	Gârbacea *et al.*	10
[2]	University of Bari	uniba	Basile *et al.*	2
[8]	University of Alberta	ualberta	Guo *et al.*	1
[9]	Amrita Vishwa Vidyapeetham	cen_neel	Barathi Ganesh *et al.*	1
[13]	IIT Kharagpur	tcs-iitkgp	Sinha *et al.*	3

Table 2: Overview summary of approaches applied in the #Microposts2015 NEEL Challenge.

Step	Method	Features	Knowledge Base	Off-the-Shelf Systems
Preprocessing	Cleaning Expansion Extraction	stop words, spelling dictionary, acronyms, hashtags, Twitter accounts, tweet timestamps, punctuation, capitalization, token positions		
Entity Mention Detection	Approximate String Matching, Exact String Matching, Fuzzy String Matching, Acronym Search Perfect String Matching, Levenshtein Matching, Jaccard String Matching, Prior Probability Matching, Context Similarity Matching, Conditional Random Fields, Random Forest	POS, tokens and adjacent tokens, contextual features, tweet timestamps, string similarity, n-grams, proper nouns, mention similarity score, Wikipedia titles, Wikipedia redirects, Wikipedia anchors, word embeddings	Wikipedia, DBpedia	Semanticizer
Entity Typing	DBpedia Type, Logistic Regression, Random Forest, Conditional Random Fields	tokens, linguistic features, word embeddings, entity mentions, NIL mentions DBpedia and Freebase types	DBpedia Freebase	
Candidate Selection	Distributional Semantic Model, Random Forest, RankSVM, Random Walk with Restart, Learning to Rank	gloss, contextual features, graph distance	Wikipedia, DBpedia	DBpedia Spotlight
NIL Detection	Conditional Random Fields, Random Forest, Lack of candidate, Score Threshold	POS, contextual words, n-grams length, predicted entity types, capitalization ratio		
NIL Clustering	Surface Form Aggregation, Type Aggregation	entity mention label, entity mention type		

entity mention (the normalization of the type is performed via a manual alignment from the DBpedia ontology and the NEEL taxonomy). The NIL is finally solved using a clustering algorithm operating on the lexical similarity of the entity mentions that do not have any DBpedia referents. To resolve the entity mention overlaps, they create a graph of all non-overlapping mentions, and assign a link score (non-linked mentions get a fixed score). They then find the highest scoring path through the graph using dynamic programming, and return the mentions of this path as the resolved list of mentions.

The system presented in Basile et al. [2] also follows a sequential workflow of mention detection and candidate selection. For the former, two approaches are built: an unsupervised based on the extraction of n-grams ($n = 0..5$), and a supervised based on the prediction of the entity boundaries from a POS tagger. Each potential entity mention is then matched with a list of DBpedia concept titles using the Levenshtein Distance, Jaccard Index, and Lucene similarity output. A filter of the entity mentions is applied with a similarity threshold of 0.85. The candidate selection stage then resolves the ambiguity of the several potential links identifying an entity mention through an adaptation of the distributional Lesk algorithm [1]. Finally, entity typing is carried out by inheriting the DBpedia type of the DBpedia reference entity pointed to, and then manually aligning this to the NEEL taxonomy.

In [8], Guo et al., present a sequential approach to the NEEL task. First, they generate potential entity mentions, using TwitIE. They then link those mentions to corresponding DBpedia referents via a candidate selection algorithm based on the similarity of the text to a dictionary built from Wikipedia titles, redirect pages, disambiguation pages and anchor text. Mentions that are not linkable are flagged as NIL. The problem of finding the correct candidate to be linked to each mention is tackled using Random Walks. Starting from the candidate links retrieved from DBpedia, a subgraph of DBpedia is built adding all adjacent entity mentions to the candidates. A personalized PageRank is then executed, giving more importance to unambiguous entities. Finally, measures of semantic relatedness between entity links, prior probability and context similarity are combined to compute an overall score. The candidate with the highest score is considered as the correct link. NIL clustering uses string similarity of entity mention names.

In [9], Barathi et al., present another sequential pipeline to the 2015 challenge, composed of generation of potential entity mentions, mention detection and candidate selection. The first stage is tackled with a linguistic approach that tokenizes the text according to Twitter cues, such as hashtags and emoticons, using the TwitIE tagger. The system then classifies entity mentions by applying a supervised learning approach using direct (e.g., POS tags) and indirect features (two words on the left and right of a candidate mention entity). In total, the authors use 34 lexical features and experiment with 3 different supervised learning algorithms. The final system implements what is determined to be the best entity recognition configuration, based on the performance achieved in the development test. The candidate selection stage is tackled by looking up DBpedia referent links. The candidate link which maximizes the similarity score between related entries and the mentions is designated as the representative. Entity mentions without related links are assigned to NIL.

Sinha et al., [13] also follow a sequential approach to the challenge task, by first detecting entity mentions from the text, and then selecting the most representative DBpedia referents (candidate selection). The first stage grounds on the linguistic cues extracted from conventional linguistic approaches such as POS tagging, word capitalization, and hashtag in the tweet. A Conditional Random Field (CRF) classifier is then trained with the linguistic features and the contextual similarity of adjacent tokens, with token window set to 5. The candidate selection is performed using an entity resolution mechanism that takes as input both the output of the entity mention detection stage and the output of DBpedia Spotlight [5]. For each entity returned from DBpedia Spotlight, if (i) the retrieved entity is found to be a substring of any of the extracted mentions in the entity mention detection stage, and if (ii) a substring match is found, then the corresponding DBpedia referent is returned and assigned to the final entity mention. If there is no match to the mention entities being extracted by the entity mention detection stage and those extracted by DBpedia Spotlight, they are assigned as NIL.

4. CORPUS CREATION AND ANNOTATION

In this section we describe the challenge dataset and the annotation process for characterising it and generating the Gold Standard. Since the challenge task was to automatically recognise, type, and link named entities (either to DBpedia referents or NIL identifiers), we built the challenge dataset considering both event and non-event tweets. While event tweets are more likely to contain named entities, non-event tweets enable us to evaluate system performance in avoiding false positives in the mention detection and candidate selection stages. The challenge dataset comprises tweets from the years 2011, 2013 and 2014. Tweets from 2011 and 2013 were extracted from a collection of over 18 million tweets provided by the Redites project.[7] These tweets cover multiple noteworthy events from 2011 and 2013 (including the death of Amy Winehouse, the London Riots, the Oslo bombing and the Westgate Shopping Mall terrorist attack). To obtain a dataset containing both event and non-event tweets, we also collected tweets from the Twitter firehose in November 2014 covering both event (such as the UCI Cyclo-cross World Cup) and non-event tweets.

Table 3: General statistics of the #Microposts2015 NEEL corpus. Dev refers to the Development set, while NEs refers to Named Entities.

	Training	Dev	Test
No. of Tweets	3,498	500	2,027
No. of Words	13,752	3,281	10,274
No. of Tokens	67,393	7,845	35,558
Avg. Tokens/Tweet	19.27	15.69	17.54
No. of Tweets with NEs	2,023	387	1,663
No. of NEs	4,016	790	3,860
No. of NIL NEs	451	362	1,478
No. of NEs with Referents	3,565	428	2,382
Avg. NEs/Tweet	1.985	2.041	2.321
Avg. NIL NEs/Tweet	0.222	0.935	0.888
Avg. NEs with Referents/Tweet	1.762	1.105	1.432

4.1 Corpus Description

The corpus consists of three main datasets: Training (58%), Development (8%) – which enabled participants to tune their systems – and Test (34%). The statistics describing the data are provided in Table 3.[8] The Training set comprises 3,498 tweets, with 67,393 tokens and 4,016 named entities. This dataset corresponds to the entire corpus of the #Microposts2014 NEEL Challenge[9] (Training + Test sets), extended with annotations for additional entity types (including Character, Event, Product, Thing) and NIL references. We also harmonized the candidate selection with the rigid designation of entity in this challenge. The Development dataset consists of 500 tweets, with 7,845 tokens and 790 named entities, while the Test set contains 35,558 tokens and 3,860 named entities. These two datasets were created by excluding the #Microposts2014 NEEL tweets from the 2015 challenge dataset, and randomly splitting the remaining tweets. The Training dataset presented a higher rate of named entities linked to DBpedia (88.76%), while the Development and Test sets were more challenging, presenting only 54.18% and 61.71% respectively. The percentage of tweets mentioning at least one entity is 57.83% in the Training set, 77.4% in the Development (Dev) set, and 82.05% in the Test set. There is very little overlap of named entities between the Training and Test data, with 4.6% (186) of the named entities in the Training also occurring in the Test set.

Summary statistics of the entity types are provided in Table 4. Across the 3 datasets the most frequent types are Person, Organization and

[7] http://demeter.inf.ed.ac.uk/redites

[8] For the computation of the statistics, the tweets were tokenized using the TwitterNLP tool (http://www.ark.cs.cmu.edu/TweetNLP).

[9] http://ceur-ws.org/Vol-1141/microposts2014-neel_challenge_gs.zip

Location. The Training dataset presents a higher rate of Organization and Thing types on average, compared to the Dev and Test datasets. The Dev dataset presents a higher rate of named entities mentioning events. The Test dataset presents a higher rate of Location. Product-types are distributed nearly evenly across the three datasets. The distributional differences between the entity types in the three sets can be clearly seen. This makes the #Microposts2015 NEEL task challenging, particularly when tackled with supervised learning approaches.

Table 4: Entity type statistics for the three data sets. Dev refers to the Development set.

Type	Training	Dev	Test
Character	43 (1.07%)	5 (0.63%)	15 (0.39%)
Event	182 (4.53%)	81 (10.25%)	219 (5.67%)
Location	786 (19.57%)	132 (16.71%)	957 (24.79%)
Organization	968 (24.10%)	125(15.82%)	541 (14.02%)
Person	1102 (27.44%)	342 (43.29%)	1402 (36.32%)
Product	541 (13.47%)	80 (10.13%)	575 (14.9%)
Thing	394 (9.81%)	25 (3.16%)	151 (3.92%)

4.2 Generating the Gold Standard

The Gold Standard (GS) was generated with the help of 3 annotators. The annotation process followed six stages.

Stage 1. Unsupervised annotation of the corpus was performed, to extract the potential entity mentions, along with the corresponding entity types and candidate links to DBpedia, that were used as input to the next stage. At this stage we used the system described in [12] for annotation.

Stage 2. The data set was divided into 3 batches (Training, Development, Test). Two annotators, using GATE,[10] annotated each batch. GATE was selected because the annotation process is guided by an ontology-centric view. However, we encountered a few issues adding the link property to each annotation, which slowed down the process, because of low flexibility in interaction with the user interface. A set of guidelines for annotation was also written, to guide the annotators in *i)* selecting the entity mentions, their types, and the corresponding candidate links provided in the first stage, and then *ii)* adding any missing annotation. The annotators were also asked to mark any problematic cases encountered.

Stage 3. A third annotator, knowledgeable about the protocol followed in Stages 1 and 2, went through the problematic cases and, involving the two initial annotators, refined the annotation procedures. The annotators then looped through stages 2 and 3 of the process till most problematic cases were resolved.

Stage 4. Unsupervised NIL Clustering generation, based on mention strings and their types, was performed.

Stage 5. The third annotator went through all NILs to include or exclude them from a given cluster. The number of mentions per NIL cluster is presented in Table 5. This shows that the Entity Type Event represented a tougher challenge for the NIL Clustering ,while the other Types had, on average, number of mentions very close to one.

Stage 6. the so-called *Adjudication Stage*, where the challenge participants reported incorrect or missing annotations. Each reported mention was evaluated by one of the challenge chairs to check compliance with the Challenge Annotation Guidelines, and additions and corrections made as required.

[10] https://gate.ac.uk

Table 5: Average number of mentions per NIL Cluster for each Named Entity type.

Type	Training	Dev	Test
Character	1.50	1.00	1.00
Event	1.67	4.50	6.11
Location	1.00	1.00	1.20
Organization	1.52	1.08	1.24
Person	1.12	1.16	1.50
Product	1.96	1.03	1.36
Thing	1.00	1.00	1.00

5. CHALLENGE RANKING

Table 6 provides the #Microposts2015 NEEL rankings. As a baseline we used a state-of-the-art approach for recognizing and linking entities from short text that is developed by *acubelab*. The system is described in [11]. The ranking is based on Equation 1, which linearly weights the contribution of the 3 metrics used in the evaluation, measuring, respectively, the contribution of the clustering approach (mention_ceaf), the typing component (strong_typed_mention_match) and the linking stage (strong_link_match). Team *ousia* [15] outperformed all other participants, with a 69% performance increase with respect to the second ranked approach, the baseline system. The top-ranked approach in this noisy context underlines current and ongoing research and industrial path in pushing toward an End-to-End system, augmented by the linguistic strength of a conventional pipeline used to filter out the irrelevant entity mentions. This approach recasts the NIL clustering stage and a supervised learning approach in predicting the role and the type of named entities that are not yet available in a Knowledge Base, such as emergent named entities, or named entities not in the scope of the Knowledge Base.

The Annotation results for the group tcs-iitkgp [13] were excluded from the ranking as they were not compatible with the challenge guidelines.

Table 6: Final #Microposts2015 NEEL Ranking

Rank	Reference	Team Name	run_{ID}	r_S
1	[15]	ousia	9	0.8067
2	**[11]**	**acubelab**	**7**	**0.4757**
3	[6]	uva	2	0.4756
4	[2]	uniba	uniba-sup	0.4329
5	[8]	ualberta	ualberta	0.3808
6	[9]	cen_neel	cen_neel_1	0.0004

Table 7 details the performance according to the *metric mention_ceaf* of the top ranked run for each participant. The runs are sorted according to the F_1 measure.

Table 7: Breakdown mention_ceaf figures per participant.

Rank	Reference	Team Name	run_{ID}	F_1
1	[15]	ousia	9	0.84
2	[6]	uva	2	0.643
3	**[11]**	**acubelab**	**7**	**0.506**
4	[2]	uniba	uniba-sup	0.459
5	[8]	ualberta	ualberta	0.394
6	[9]	cen_neel	cen_neel_1	0.001

Table 8 reports the performance of the top ranked run per participant according to the metric *strong_typed_mention_match*. The runs are sorted according to the F_1 measure.

Table 8: Breakdown strong_typed_mention_match figures per participant.

Rank	Reference	Team Name	run_{ID}	F_1
1	[15]	ousia	9	0.807
2	[6]	uva	2	0.412
3	**[11]**	**acubelab**	**7**	**0.388**
4	[2]	uniba	uniba-sup	0.367
5	[8]	ualberta	ualberta	0.329
6	[9]	cen_neel	cen_neel_1	0

Table 9 reports the performance of the top ranked run per participant according to the metric *strong_link_match*. The runs are sorted according to the F_1 measure.

Table 9: Breakdown strong_link_match figures per participant.

Rank	Reference	Team Name	run_{ID}	F_1
1	[15]	ousia	9	0.0.762
2	**[11]**	**acubelab**	**7**	**0.523**
3	[2]	uniba	uniba-sup	0.464
4	[8]	ualberta	ualberta	0.415
5	[6]	uva	2	0.316
6	[9]	cen_neel	cen_neel_1	0

Table 10 reports the performance of the top ranked run per participant based on latency (expressed in seconds s). Each measure is reported along with the confidence interval obtained from the selection procedure of the annotation results as reported in 2.2.2.

Finally, Table 11 shows the breakdown for the best 3 runs per participant over all metrics used in the evaluation of the systems.

6. CONCLUSIONS

The #Microposts2014 NEEL challenge was to foster the development of novel approaches for entity extraction, and linking in Microposts. In 2015 the NEEL task was extended to include integration of named entity typing and the characterization of entities to either DBpedia referents or NIL references. The motivation for organizing this challenge is the strong, current interest of the research and commercial communities in developing systems able to fit the challenging context of Microposts in entity extraction, entity recognition, and entity linking. Although state-of-the-art approaches offer a large number of options for tackling the challenge

Table 10: Breakdown latency figures per participant.

Rank	Reference	Team Name	run_{ID}	$[s]$
1	**[11]**	**acubelab**	**7**	**0.13±0.02**
2	[6]	uva	2	0.19±0.09
3	[2]	uniba	uniba-sup	2.03±2.35
4	[8]	ualberta	ualberta	3.41±7.62
3	[15]	ousia	9	8.5±3.62
6	[9]	cen_neel	cen_neel_1	12.37±27.6

task, the evaluation results show that the NEEL task remains challenging when applied to tweets with their peculiarities, compared to standard, lengthy texts.

The evaluation strategy used in the 2014 challenge has been extended in 2015, to account for *mention_ceaf*, *strong_link_match*, *strong_typed_mention_match* and *latency*, following the established metrics introduced in the TAC KBP 2014 task. Carrying out evaluation in this way provided a more robust approach for ranking participants' entries.

As a result of the 2015 NEEL challenge we have generated a manually annotated corpus, which extends that in 2014 with the annotation of typed entities and the generation of NIL identifiers. To the best of our knowledge this is the largest publicly available corpus providing named entities, types, and link annotations for Microposts. The gold standard[11] is released with the CC BY 4.0 license.[12] We hope that through our release of data and resources, we will promote research on entity recognition and disambiguation, especially with regard to Microposts.

Our evaluation results report a clear winner: Team *ousia* [15] consolidated and, further, extended the findings of the NEEL 2014 winner, using an End-to-End system for both candidate selection and mention typing, along with a linguistic pipeline to perform entity typing and filtering.

The #Microposts2015 NEEL challenge saw a considerable drop in participants after the initial intent to participate. Among the participants who withdrew, reasons given were mainly poor results from their prototypes and the complexity in developing a reliable prototype to be deployed as a Web service. Aiming to consolidate the current challenge task we believe it will aid participants in such challenges to further develop their prototypes by providing a base engineering platform for deployment in a live context. We have, in 2015, also built bridges with the TAC community. We plan to strengthen these and to involve a larger audience of potential participants spanning the Linguistics, Machine Learning, Knowledge Extraction and Data Semantics fields, in order to widen the scope for potential solutions to what is acknowledged to be a challenging, albeit valuable, exercise.

7. ACKNOWLEDGMENTS

Especial thanks to SpazioDati who generously sponsored the prize for the winning submission. We thank also the participants who helped to improve the Gold Standard.
Giuseppe Rizzo is funded by LinkedTV (GA No. 287911), Amparo

[11] http://ceur-ws.org/Vol-1395/microposts2015_neel-challenge-report/microposts2015-neel_challenge_gs.zip
[12] http://creativecommons.org/licenses/by/4.0

Table 11: Top 3 runs per participant, sorted according to r_S.

Rank	Reference	Team Name	run_{ID}	$tagging_{F1}$	$clustering_{F1}$	$linking_{F1}$	$latency[s]$	r_S
1	[15]	ousia	9	0.807	0.84	0.762	8.5±3.62	0.8067
2	[15]	ousia	5	0.68	0.843	0.762	8.48±3.6	0.7698
3	[15]	ousia	10	0.679	0.842	0.762	8.49±3.57	0.7691
4	[11]	acubelab	7	0.388	0.506	0.523	0.13±0.02	0.4757
5	[6]	uva	2	0.412	0.643	0.316	0.19±0.09	0.4756
6	[11]	acubelab	6	0.385	0.506	0.524	0.13±0.02	0.4751
7	[11]	acubelab	9	0.388	0.506	0.523	0.13±0.02	0.4734
8	[6]	uva	3	0.404	0.642	0.285	0.19±0.1	0.4635
9	[6]	uva	6	0.383	0.595	0.318	1.73±0.86	0.4483
10	[2]	uniba	uniba-sup	0.367	0.459	0.464	2.03±2.35	0.4329
11	[8]	ualberta	ualberta	0.329	0.394	0.415	3.41±7.62	0.3808
12	[2]	uniba	uniba-unsup	0.367	0.459	0.464	2.03±2.35	0.4329
13	[9]	cen_neel	cen_neel_1	0	0.001	0	12.89±27.6	0.004

E. Cano is funded by the MK:Smart project and Bianca Pereira by the Science Foundation Ireland (GA No. SFI/12/RC/2289).

8. REFERENCES

[1] P. Basile, A. Caputo, and G. Semeraro. An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. In *25th International Conference on Computational Linguistics (COLING'14)*, 2014.

[2] P. Basile, A. Caputo, G. Semeraro, and F. Narducci. UNIBA: Exploiting a distributional semantic model for disambiguating and linking entities in tweets. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 62–63, 2015.

[3] A. E. Cano Basave, G. Rizzo, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. Making Sense of Microposts (#Microposts2014) Named Entity Extraction & Linking Challenge. In *4th Workshop on Making Sense of Microposts (#Microposts2014)*, 2014.

[4] A. E. Cano Basave, A. Varga, M. Rowe, M. Stankovic, and A.-S. Dadzie. Making Sense of Microposts (#msm2013) Concept Extraction Challenge. In *Making Sense of Microposts (#MSM2013) Concept Extraction Challenge*, 2013.

[5] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *9th International Conference on Semantic Systems (I-SEMANTICS '13)*, 2013.

[6] C. Gârbacea, D. Odijk, D. Graus, I. Sijaranamual, and M. de Rijke. Combining multiple signals for semanticizing tweets: University of Amsterdam at #Microposts2015. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 59–60, 2015.

[7] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'11)*, 2011.

[8] Z. Guo and D. Barbosa. Entity recognition and linking on tweets with random walks. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 57–58, 2015.

[9] B. G. H B, A. N, A. K. M, V. R, and S. K P. AMRITA – CEN@NEEL: Identification and linking of Twitter entities. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 64–65, 2015.

[10] X. Luo. On coreference resolution performance metrics. In *Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP'05)*, 2005.

[11] F. Piccinno and P. Ferragina. From TagME to WAT: A New Entity Annotator. In *1st International Workshop on Entity Recognition & Disambiguation (ERD '14)*, 2014.

[12] G. Rizzo, M. van Erp, and R. Troncy. Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web. In *9th International Conference on Language Resources and Evaluation (LREC'14)*, 2014.

[13] P. Sinha and B. Barik. Named entity extraction and linking in #Microposts. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 66–67, 2015.

[14] R. Walpole and R. Myers. *Probability and statistics for engineers & scientists (Eighth Edition)*. Pearson Education International, 2007.

[15] I. Yamada, H. Takeda, and Y. Takefuji. An end-to-end entity linking approach for tweets. In *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 55–56, 2015.

APPENDIX

A. NEEL TAXONOMY

`Thing`
- languages
- ethnic groups
- nationalities
- religions
- diseases
- sports
- astronomical objects

Examples:
```
If all the #[Sagittarius] in the world
Jon Hamm is an [American] actor
```

`Event`
- holidays
- sport events
- political events

social events

Examples:
```
[London Riots]
[2nd World War]
[Tour de France]
[Christmas]
[Thanksgiving] occurs the ...
```

`Character`
fictional characters
comic characters
title characters

Examples:
```
[Batman]
[Wolverine]
[Donald Draper]
[Harry Potter] is the strongest wizard in
the school
```

`Location`
public places (squares, opera houses, museums, schools, markets, airports, stations, swimming pools, hospitals, sports facilities, youth centers, parks, town halls, theatres, cinemas, galleries, universities, churches, medical centers, parking lots, cemeteries)
regions (villages, towns, cities, provinces, countries, continents, dioceses, parishes) commercial places (pubs, restaurants, depots, hostels, hotels, industrial parks, nightclubs, music venues, bike shops)
buildings (houses, monasteries, creches, mills, army barracks, castles, retirement homes, towers, halls, rooms, vicarages, courtyards)

Examples:
```
[Miami]
Paul McCartney at [Yankee Stadium]
president of [united states]
Five New [Apple Retail Store] Opening
Around
```

`Organization`
companies (press agencies, studios, banks, stock markets, manufacturers, cooperatives)
subdivisions of companies
brands
political parties
government bodies (ministries, councils, courts, political unions)
press names (magazines, newspapers, journals)
public organizations (schools, universities, charities)
collections of people (sport teams, associations, theater companies, religious orders, youth organizations, musical bands)

Examples:
```
[Apple] has updated Mac Os X
[Celtics] won against
[Police] intervene after disturbances
[Prism] performed in Washington
[US] has beaten the Japanese team
```

`Person`
people's names (titles and roles are not included, such as Dr. or President)

Examples:
```
[Barack Obama] is the current
[Jon Hamm] is an American actor
[Paul McCartney] at Yankee Stadium
call it [Lady Gaga]
```

`Product`
movies
tv series
music albums
press products (journals, newspapers, magazines, books, blogs)
devices (cars, vehicles, electronic devices)
operating systems
programming languages

Examples:
```
Apple has updated [Mac Os X]
Big crowd at the [Today Show]
[Harry Potter] has beaten any records
Washington's program [Prism]
```

Section IVa:

NEEL Challenge Submissions I

An End-to-End Entity Linking Approach for Tweets

Ikuya Yamada[1 2 3]
ikuya@ousia.jp

Hideaki Takeda[3]
takeda@nii.ac.jp

Yoshiyasu Takefuji[2]
takefuji@sfc.keio.ac.jp

[1]Studio Ousia Inc., 4489-105-221 Endo, Fujisawa, Kanagawa, Japan
[2]Keio University, 5322 Endo, Fujisawa, Kanagawa, Japan
[3]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan

ABSTRACT

We present a novel approach for detecting, classifying, and linking entities from Twitter posts (tweets). The task is challenging because of the *noisy*, *short*, and *informal* nature of tweets. Consequently, the proposed approach introduces several methods that robustly facilitate successful realization of the task with enhanced performance in several measures.

Keywords

Entity linking; Wikification; Twitter; DBpedia; Wikipedia

1. INTRODUCTION

Microblogging services, such as *Twitter*, are rapidly becoming virtually ubiquitous. This is attributable to the fact that they are extremely valuable mechanisms that enable us to obtain live and raw information in real time. In this paper, we describe our approach to the #Microposts 2015 NEEL challenge [6], a competition for extracting and typing entity mentions appearing in tweets, and linking those mentions to the corresponding URIs of the DBpedia 2014 dataset[1], with non-existent mentions also being recognized as *NIL* mentions.

The main difficulty inherent in this task stems from the *noisy*, *short*, and *informal* nature of tweets. The performance of previous approaches suffered because they tended to focus on well-written, long texts such as news articles. Our system explicitly focuses on tweets and addresses the problem using a variety of methods working together.

Our proposed system addresses the task in an *end-to-end* manner. Unlike most of the previous approaches, the system does not use an external named entity recognition system (NER) to generate candidates of the entity mentions because the current NER typically performs badly for tweets [5]. Our system first generates the candidates by using *approximate candidate generation* that can detect *misspelled* and *abbreviated* mentions and *acronyms*. Then it uses *supervised machine-learning* to remove irrelevant candidates and resolve them into the corresponding DBpedia URIs.

[1]`http://wiki.dbpedia.org/Downloads2014`

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (`http://ceur-ws.org/Vol-1395`)

#Microposts2015, May 18th, 2015, Florence, Italy.

Consequently, we constructed three supervised machine-learning models to detect NIL entity mentions and predict the types (e.g., *PERSON* and *LOCATION*) of the detected mentions.

2. THE PROPOSED SYSTEM

Our proposed system addresses the task using a procedure comprising the following five steps: 1) preprocessing, 2) mention candidate generation, 3) mention detection and disambiguation, 4) NIL mention detection, and 5) type prediction.

2.1 Preprocessing

We tokenize a tweet and assign part-of-speech tags to the resulting tokens using ARK Twitter Part-of-Speech Tagger [2] with our enhanced hashtag tokenization method. We also extract the timestamp of the tweet from the Tweet ID.

2.2 Mention Candidate Generation

In this step, the candidates of the entity mentions are generated from the tweet using the methods described below.

Mention-Entity Dictionary.

The system uses a *mention-entity* dictionary that maps mention surface (e.g., *apple*) to the possible referent entities (e.g., Apple_Inc., Apple (food)). The possible mention surfaces of an entity are extracted from the corresponding Wikipedia page title, the page titles of the Wikipedia pages that redirect to the page of the entity, and anchor texts in Wikipedia articles that point to the page of the entity. We constructed this dictionary using the January 2015 dump of Wikipedia.

Candidate Generation Methods.

The system generates candidates using the mention-entity dictionary; it first takes all the n-grams ($n < 10$) from the tweet and performs queries to the dictionary using the text surface of each of these n-grams. The following four methods are used to retrieve candidates:

- *Exact search* retrieves mention candidates that have text surfaces exactly equal to the query text.
- *Fuzzy match* searches the mention candidates that have text surfaces within a certain distance of the query text measured by edit distance.
- *Approximate token search* obtains mention candidates whose text surfaces have a significant ratio of words in common with the query text.
- *Acronym search* retrieves mention candidates with possible acronyms[2] that include the query text.

[2]We generate acronyms by tokenizing the mention surface and simply taking the first characters of the resulting tokens.

The system first generates possible mention candidates using the above methods, sorts these candidates according to the number of occurrences in which the mention appear as a link to the referent entity, and selects the top k candidates ($k = 100$ for exact search and $k = 30$ for other methods). Additionally, we experimentally set the maximum allowed edit distance of *fuzzy match* to *two* and the minimum ratio of *approximate token search* to *66%* because these settings achieve the best scores in our experiments.

2.3 Mention Detection and Disambiguation

In this step, we first assign a score to mention candidates using a supervised machine-learning model. In this case, we used *random forest* as the machine-learning algorithm.

Features.

We started out using features similar to those proposed in previous works [1, 3], and subsequently introduced several novel features to enhance performance. The features introduced include 1) *contextual information using word embeddings* to measure the contextual similarity between a tweet and an entity, 2) *temporal popularity knowledge of an entity* extracted from Wikipedia page view data, and 3) *string similarity measures* to measure the similarity between the title of the entity and the mention (e.g., edit distance).

Overlap Resolution.

Finally, the overlapped entity mentions are resolved. We start with the beginning of the tweet and iterate over the candidate entity mentions. Then, we detect the mention if the corresponding span of the mention has not already been detected and the score assigned to the mention is above the threshold. If multiple mentions are found, the mention with the highest score is selected.

2.4 NIL Mention Detection

We formulate the task of detecting NIL mentions from a tweet as a supervised classification task to assign a binary label to each of all possible n-grams ($n < 10$). *Random forest* is again used as our machine-learning algorithm.

Features.

We extract several features from the output of the Stanford NER[3] using two types of models: 1) a standard three-class model, and 2) a model that does not use capitalization as a feature. We also use the ratio of capitalized words as an indicator of the reliability of the capitalization in the tweet. Additionally, various other features are used, such as part-of-speech tags of the surrounding words and the length of the n-grams.

2.5 Type Prediction

We cast the task of detecting types of mentions as a multi-class supervised classification task. In the previous steps, we extracted two types of mentions: entity mentions and NIL mentions. Thus, we are able to build two separate classifiers to predict the entity types for each type of mention. We developed two machine-learning models using *logistic regression* and *random forest* and created the final model by building an ensemble model on top of these models in order to boost the performance.

Features for Entity Mentions.

The primary features used to detect types of entity mentions are the corresponding entity classes retrieved from *DBpedia* and *Freebase* (e.g., FictionalCharacter, SportsTeam). We also use our 300 dimensional entity-embeddings constructed from Wikipedia and the predicted entity types of the Stanford NER.

[3] `http://nlp.stanford.edu/software/CRF-NER.shtml`

Name	Precision	Recall	F1
strong_link_match	0.786	0.656	0.715
strong_typed_mention_match	0.656	0.630	0.642
mention_ceaf	0.857	0.823	0.840

Table 1: Summary of experimental results

Features for NIL Mentions.

In order to detect the types of NIL mentions, we use features extracted from word embeddings. Here, the GloVe Twitter 2B model [4] is used as the word embeddings. We also use the predicted types of the Stanford NER and the part-of-speech tags.

3. EXPERIMENTAL RESULTS

In our experiments, we used the #Microposts 2015 dataset [6] split into a training set and a test set. These sets contained *3,498* and *500* tweets respectively.

Table 1 shows a summary of our experimental results. We evaluated our system using the following three measures: *strong_link_match* to evaluate the performance of linking entities, *strong_typed_mention_match* to measure the performance of mention detection and entity typing, and *mention_ceaf* for calculating the performance of clustering detected mentions into entity mentions or NIL mentions.[4] We successfully achieved accurate performance in all of the measures.

4. CONCLUSIONS

In this paper, we described our approach for detecting, classifying, and linking entity mentions in tweets. We introduced a novel machine-learning approach specifically targeted at tweets and successfully achieved enhanced performance on the #Microposts2015 dataset.

References

[1] P. Ferragina and U. Scaiella. TAGME: On-the-fly Annotation of Short Text Fragments (by Wikipedia Entities). In *CIKM '10*, pages 1625–1628, 2010.

[2] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *ACL '11*, pages 42–47, 2011.

[3] E. Meij, W. Weerkamp, and M. de Rijke. Adding Semantics to Microblog Posts. In *WSDM '12*, pages 563–572, 2012.

[4] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In *EMNLP '14*, pages 1532–1543, 2014.

[5] A. Ritter, S. Clark, and O. Etzioni. Named Entity Recognition in Tweets: An Experimental Study. In *EMNLP '11*, pages 1524–1534, 2011.

[6] G. Rizzo, A. E. Cano Basave, B. Pereira, and A. Varga. Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge. In M. Rowe, M. Stankovic, and A.-S. Dadzie, editors, *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 44–53, 2015.

[4] For further details of these measures, please refer to `https://github.com/wikilinks/neleval/wiki/Evaluation`

Entity Recognition and Linking on Tweets with Random Walks

Zhaochen Guo
Department of Computing Science
University of Alberta
zhaochen@ualberta.ca

Denilson Barbosa
Department of Computing Science
University of Alberta
denilson@ualberta.ca

ABSTRACT

This paper presents our system at the #Microposts2015 NEEL Challenge [4]. The task is to recognize and type mentions from English Microposts, and link them to their corresponding entries in DBpedia 2014. For this task, we developed a method based on a state-of-the-art entity linking system - REL-RW [2], which exploits the entity graph from the knowledge base to compute semantic relatedness between entities, and use it for entity disambiguation. The advantage of the approach is its robustness for various types of documents. We built our system on REL-RW and employed a tweet specific NER component to improve the performance on tweets. The system achieved overall 0.35 F1 on the development dataset from NEEL 2015, while the disambiguation component alone can achieve 0.70 F1.

Keywords

Entity Recognition, Entity Disambiguation, Social Media

1. INTRODUCTION

Microposts such as tweets become popular nowadays. The tweets, though short and simple, can spread information fast and broadly. Events, reviews, news and so on are all posted on Twitter, which make tweets a very valuable resource to support many activities such as political option mining, product development (customer review), or social activism. We need to understand the tweets to make best use of them for such applications. Given the maximum 140 characters limit, there is barely enough useful information in a tweet. Exploiting entities mentioned in tweets can enrich the text with their contexts and semantics in knowledge bases, which is important for a better understanding of tweets. The NEEL task aims to solve this issue by automatically recognizing entities and their types from English tweets, and linking them to their DBpedia 2014 resources. NER and entity linking have been active research subjects. However, most previous works focus on traditional long documents, which do not pose the challenges in tweets, such as the noisy terms, hashtags, retweets, abbreviations, and cyber-slang. Appropriately addressing these problems, and taking advantage of the existing approaches are important.

We developed a NEEL system for the challenge based on a state-of-the-art entity linking approach, and incorporated a tweet specific mention extraction component. Our system takes advantage of the entity graph in the knowledge base, and does not rely on the lexical features in the tweets, which makes it robust on different datasets. In the following sections, we will describe our system and report the experimental results on the challenge benchmarks.

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

2. OUR APPROACH

2.1 Mention Extraction

As the first component of our system, mention extraction extracts named entities from the given tweets. Our system originally employed the Stanford NER with models trained on the well-formed news documents. However, it cannot handle the short tweets very well. We then used the TwitIE [1] from GATE, a NER tool designed specifically for tweets, to perform the mention extraction in our system.

Compared to the Stanford NER, TwitIE added several improvements. The first is the *Normaliser*. To address unseen tokens and noisy grammars in tweets, TwitIE used a spelling dictionary specific to the tweets to identify and correct spelling variations. The second improvement is a tweet adapted model for the POS tagging. While still employing the Stanford POS Tagger, TwitIE replaced the original model with a new model trained on Twitter datasets which were annotated with the Penn TreeBank with extra tag labels such as retweets, URLs, hashtags and user mentions. With these improvements, TwitIE helps improve the NER performance of our system. Note that we use the types inferred from TwitIE as the types for mentions.

2.2 Candidate Generation

The second component is the candidate generation which selects potential candidates from the knowledge base for mentions in the tweets. Our system utilized an alias dictionary collected from Wikipedia titles, redirect pages, disambiguation pages, and the anchor text in Wikilinks [2], which maps each alias to entities it refers to in Wikipedia.

We simply use exact string matching against the dictionary for the candidate generation. Mentions that do not match any alias in the dictionary are immediately linked to NIL. Otherwise, the mapping entities of the matched alias are selected as candidates. To improve the efficiency, we

further prune the candidates by two criteria [2]: *prior probability* which is defined as the probability the alias refers to an entity in the Wikipedia corpus, and *context similarity* which measures the context similarity (cosine similarity) of the mention and the entity. For both criteria, the top 10 ranked candidates are selected and then merged to generate the final candidate list for the given mention.

2.3 Entity Disambiguation

Entity disambiguation is to select the target entity from the candidates of a mention. We use our prior algorithm [2] for this task. The main idea is to represent the semantics of the document (tweet) and candidate entities using a set of related entities in DBpedia for which the weight of each entity is measured by their semantic relatedness with the candidates. We then use the semantic representation to compute the semantic similarity between the candidates and the document. For each mention-entity pair, we measure their prior probability, context similarity, and semantic similarity and linearly combine them together to compute an overall similarity. The candidate with the highest similarity will be selected as the target entity.

The key part of the approach is the semantic representation and relatedness. Knowledge bases such as DBpedia are graphs where entities are connected semantically. We construct an entity graph from the knowledge base and use the connectivity in the graph to measure the semantic relatedness between entities. We use random walks with restart to traverse the graph. Upon convergence, this process results in a probability distribution over the vertices corresponding to the likelihood these vertices are visited. This probability can then be used as an estimation of relatedness between entities in the graph. For each target entity, we restart from that entity in each random walk, generating a personalized probability for the target entity, and use it as the semantic representation. For the semantic representation of the document, we perform the random walk restarting from a set of entities representing the document. Since the true entities of mentions in the documents are not available, we either choose the representative entities from the unambiguous mentions which have only one candidate, or the candidate entities whose weights are approximated by their prior probability. With the representative entities, the semantic representation of the document can then be computed as the probability distribution obtained through the random walk from these entities.

To improve the efficiency, instead of using the entire DBpedia graph, we construct a small entity graph by starting with the set of candidates, and adding all entities adjacent to these candidates in the original graph. This subgraph contains entities semantically related to the candidates and is large enough to compute the semantic representation of entities and the document.

Once obtaining the semantic representation, we measure the semantic similarity between each candidate and the document using the Zero-KL Divergence [3], which is then combined with the prior probability and context similarity to disambiguate candidates.

2.4 NIL Prediction and Clustering

For NIL prediction, mentions are deemed out of a knowledge base (and thus linked to NIL) either when no candidates are available or their similarity with the highest ranked entities is below a threshold. For clustering, we simply group mentions by their name similarity. In the future, we plan to exploit the semantic representation of the tweets to measure their semantic similarity and use it for NIL clustering.

	Precision	Recall	F1
Tagging	0.34	0.22	0.27
Linking	0.35	0.36	0.35
Clustering	0.45	0.29	0.35

Table 1: Results on the development datasets.

3. EXPERIMENTS

We built our system using a 2013 DBpedia dump, including the knowledge base and alias dictionary. Table 3 lists the results of our system on the development dataset. As shown, the performance of the mention extraction (tagging) is very poor, especially the recall. We believe more tuning would improve the performance. Since the novelty of our system is the disambiguation part, we further evaluated the performance of the entity disambiguation component separately (assuming all mentions are correctly recognised), and the system can achieve results of 0.74 precision, 0.66 recall for an F1 of 0.70 on the dataset.

4. CONCLUSION

In this paper, we described a system for the #Microp-ost2015 NEEL challenge, in which we adopted a tweet specific NER system for mention extraction, and used an entity disambiguation approach that utilized the connectivity of entities in DBpedia to capture the semantics of entities and disambiguate mentions.

Due to time limitation, our system still has much room for improvements. As shown, mention extraction is now the bottleneck of the system and needs further improvement. More features from the tweets could be used to train a better model. For the mention disambiguation, we will explore supervised approaches such as learning to rank to combine the semantic features such as the semantic similarity and lexical features specific to tweets. Also, the semantic representation seems to be valuable for the NIL clustering and worth exploration.

5. REFERENCES

[1] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard, and N. Aswani. TwitIE: An open-source information extraction pipeline for microblog text. In *RANLP*. ACL, 2013.

[2] Z. Guo and D. Barbosa. Robust entity linking via random walks. In *CIKM*, pages 499–508, 2014.

[3] T. Hughes and D. Ramage. Lexical semantic relatedness with random graph walks. In *EMNLP-CoNLL*, pages 581–589, 2007.

[4] G. Rizzo, A. E. Cano Basave, B. Pereira, and A. Varga. Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge. In M. Rowe, M. Stankovic, and A.-S. Dadzie, editors, *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 44–53, 2015.

Combining Multiple Signals for Semanticizing Tweets: University of Amsterdam at #Microposts2015

Cristina Gârbacea, Daan Odijk, David Graus, Isaac Sijaranamual, Maarten de Rijke

University of Amsterdam, Science Park 904, Amsterdam, The Netherlands
{G.C.Garbacea, D.Odijk, D.P.Graus, I.B.Sijaranamual, deRijke}@uva.nl

ABSTRACT

In this paper we present an approach for extracting and linking entities from short and noisy microblog posts. We describe a diverse set of approaches based on the Semanticizer, an open-source entity linking framework developed at the University of Amsterdam, adapted to the task of the #Microposts2015 challenge. We consider alternatives for dealing with ambiguity that can help in the named entity extraction and linking processes. We retrieve entity candidates from multiple sources and process them in a four-step pipeline. Results show that we correctly manage to identify entity mentions (our best run attains an F1 score of 0.809 in terms of the strong mention match metric), but subsequent steps prove to be more challenging for our approach.

Keywords

Named entity extraction; Named entity linking; Social media

1. INTRODUCTION

This paper describes our participation in the named entity extraction and linking challenge at #Microposts2015. Information extraction from microblog posts is an emerging research area which presents a series of problems for the natural language processing community due to the shortness, informality and noisy lexical nature of the content. Extracting entities from tweets is a complex process typically performed in a sequential fashion. As a first step, *named entity recognition (NER)* aims to detect mentions that refer to entities, e.g., names of people, locations, organizations or products (also known as *entity detection*), and subsequently to classify the mentions into predefined categories (*entity typing*). After NER, *named entity linking (NEL)* is performed: linking the identified mentions to entries in a knowledge base (KB). Due to its richness in semantic content and coverage, Wikipedia is a commonly used KB for linking mentions to entities, or deciding when a mention refers to an entity that is not in the KB, in which case it is referenced by a NIL identifier. DBpedia aims to extract structured information from Wikipedia, and combines this information into a huge, cross-domain knowledge graph which provides explicit structure between concepts and the relations among them.

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

Our participation in this challenge revolves around the existing open-source entity linking software developed at the University of Amsterdam. We use *Semanticizer*[1], a state-of-the-art entity linking framework. So far Semanticizer has been successfully employed in linking entities in search engine queries [1] and in linking entities in short documents in streaming scenarios [6]. Moreover, it has been further extended to deal with additional types of data like television subtitles [3]. In what follows we explain how we use Semanticizer for the task at hand, and describe each of our submitted runs to the competition.

2. SYSTEM ARCHITECTURE

Our system processes each incoming tweet in four stages: mention detection, entity disambiguation and typing, NIL identification and clustering, and overlap resolution. We explain each stage in turn.

Mention detection: The first step aims to identify all entity mentions in the input text, and is oriented towards high recall. We take the union of the output of two mention identification methods:

Semanticizer: the state-of-the-art system performs lexical matching of entities' surface forms. These surface forms are derived from the KB, and comprise anchor texts that refer to Wikipedia pages, disambiguation and redirect pages, and page titles as described in Table 1. For this, we use two instances of Semanticizer, running on two Wikipedia dumps: one dated May 2014 (the version used to build DBpedia 3.9), and a more recent one, dated February 2015.

We perform three separate preprocessing steps on the tweet text, the results of which get sent to the Semanticizer. These steps are: *i)* the raw text, *ii)* the cleaned text (replacing @-mentions with corresponding Twitter account names, and splitting hashtags using dynamic programming), and *iii)* the normalized text (e.g., case-folding, removing diacritics).

NER: For identifying entity mentions that do not exist in Wikipedia, i.e., out of KB entities, we employ a state-of-the-art named entity recognizer, previously applied to finding mentions of emerging entities on Twitter [2]. We train five different NER models, three using the ground truth data from the Microposts challenges (2013 through 2015), one using pseudo-ground truth (generated by linking tweets as in [2]) and one trained on all data.

Given the candidate mentions identified by NER and Semanticizer, we include a binary feature to express whether the mention has been detected by both systems. For each mention we end up with the set of features described in Table 1 that we use in training a Random Forest classifier (using 100 trees and rebalancing the classes per tweet by modifying instance weights), to predict whether a candidate mention is an entity mention (actually refers to an entity).

[1] https://github.com/semanticize/semanticizer

Table 1: Features used for mention detection.

Feature	Description
linkOccCount	no. of times mention appears as anchor text on Wikipedia
linkDocCount	no. of docs in which mention appears as anchor text
occCount	no. of times the mention appears on Wikipedia
senseOccCount	no. of times the mention is anchor to Wikipedia title
senseDocCount	no. of docs the mention is anchor to Wikipedia title
priorProbability	% of docs where anchor links to target Wikipedia title
linkProbability	% of docs where mention is anchor for a Wikipedia link
senseProbability	% of docs where mention links to target Wikipedia article
isCommon	the mention is found by both NER and Semanticizer

Entity disambiguation and typing: Given the entity mentions from the previous stage, the next step is to identify referenced entities. We retrieve the full list of candidate entities, extract features, and cast the disambiguation step of identifying the correct entity for a mention as a learning to rank problem.

Next to the features in Table 1, we use additional full-text search features. We index Wikipedia using ElasticSearch (ES), and issue the tweet as a query for candidate entities' retrieval scores. We also retrieve the 10 most similar entities for each candidate, using a *more like this* query. Finally, we incorporate Wikipedia page view statistics[2] from April 2014 as features. We use these features to train RankSVM to rank the entity candidates for each mention, and take the top ranked candidate as the entity to link. We map the entity to its DBpedia URI, and determine its type through a manual mapping of DBpedia classes to the #Microposts2015 taxonomy.

NIL identification and clustering: To decide whether the top-ranked entity is correct, or the mention refers to an out-of-KB entity, we compute meta-features based on the RankSVM classifier's scores. We use these meta-features to train a Random Forest classifier for NIL detection. We cluster NILs by linking identical mentions to a single NIL identifier based on their surface forms.

Overlap resolution: Finally, we resolve all overlapping mentions that are output by the mention identification step. We create a graph of all non-overlapping mentions, and assign them their link score (non-linked mentions get a fixed score). We then find the highest scoring path through the graph using dynamic programming, and return the mentions of this path as our resolved list of mentions.

Our submitted runs rely on this scheme and variations thereof. See Table 2 for an overview of the runs. We hypothesize that the Semanticizer will yield high entity recall, but low precision. Filtering the resulting candidates by *senseProbability* will increase precision. We expect the NER runs to be superior to Semanticizer or ES-only runs. Finally, we believe that combining the NER and Semanticizer outputs with additional candidates returned by ES will outperform all our other runs.

3. RESULTS

We evaluate our approach on the dev set consisting of 500 tweets made available by the organizers [4], [5]. In Table 3 we report on the official metrics for entity detection, tagging, clustering and linking. Our best performing runs (Run 1, Run 2) in terms of mention detection and typing rely mainly on NER and ES features. Even though Semanticizer detects candidates with high recall, our analysis indicates that most errors occur when the system fails to recognize mentions correctly, which negatively impacts the linking scores. Since each step in the pipeline relies on the output from the previous step, cascading errors influence our results, and we believe a more in-depth error analysis of each stage is desirable. Despite its simplicity, our clustering approach performs reasonably well.

[2] https://dumps.wikimedia.org/other/pagecounts-raw/

Table 2: Description of our runs.

RunID	NER	Semanticizer	Disambiguation	Filter
Run 1	2015	-	-	-
Run 2	2015	-	full-text search	-
Run 3	2015	-	full-text search	NIL
Run 4	all	-	full-text search	-
Run 5	-	2014	senseProbability	-
Run 6	*Same as Run 5 without overlap resolution.*			
Run 7	all	all	full-text search	NIL
Run 8	2015	all	RankSVM	NIL
Run 9	all	all	RankSVM	NIL
Run 10	*Same as Run 9 with a lower mention detection threshold.*			

Table 3: F1 scores on the dev set for strong mention match (SMM), strong typed mention match (STMM), strong link match (SLM), and mention ceaf (MC) metrics.

RunID	SMM	STMM	SLM	MC
Run 1	**0.809**	0.456	0.164	0.715
Run 2	**0.809**	**0.460**	0.330	**0.731**
Run 3	**0.809**	0.455	0.291	0.730
Run 4	0.554	0.311	0.213	0.497
Run 5	0.411	0.288	0.280	0.374
Run 6	0.620	0.389	0.280	0.567
Run 7	0.533	0.330	0.210	0.486
Run 8	0.732	0.418	**0.334**	0.633
Run 9	0.577	0.365	0.247	0.525
Run 10	0.566	0.355	0.280	0.515

4. CONCLUSION

We have presented a system that performs entity mention detection, disambiguation and clustering on short and noisy text by drawing candidates from multiple sources and combining them. We observe that our simple NER and ES runs perform better than our more complex runs. We believe that more robust methods are needed to deal with the errors introduced at each step of the pipeline. For future work we plan on improving mention detection with additional Semanticizer features.

Acknowledgements. This research was partially supported by the Netherlands Organisation for Scientific Research (NWO) under project numbers 727.011.005, SEED and 640.006.013, DADAISM; Amsterdam Data Science, and the Dutch national program COMMIT.

REFERENCES

[1] D. Graus, D. Odijk, M. Tsagkias, W. Weerkamp, and M. de Rijke. Semanticizing search engine queries: the University of Amsterdam at the ERD 2014 challenge. In *The first international workshop on Entity recognition & disambiguation*, 2014.

[2] D. Graus, M. Tsagkias, L. Buitinck, and M. de Rijke. Generating pseudo-ground truth for predicting new concepts in social streams. In *ECIR 2014*. Springer, 2014.

[3] D. Odijk, E. Meij, and M. de Rijke. Feeding the second screen: Semantic linking based on subtitles. In *OAIR 2013*, 2013.

[4] G. Rizzo, A. E. Cano Basave, B. Pereira, and A. Varga. Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge. In Rowe et al. [5], pages 44–53.

[5] M. Rowe, M. Stankovic, and A.-S. Dadzie, editors. *Proceedings, 5th Workshop on Making Sense of Microposts (#Microposts2015): Big things come in small packages, Florence, Italy, 18th of May 2015*, 2015.

[6] N. Voskarides, D. Odijk, M. Tsagkias, W. Weerkamp, and M. de Rijke. Query-dependent contextualization of streaming data. In *ECIR 2014*. Springer, 2014.

Section IVb:

NEEL Challenge Submissions II Posters

UNIBA: Exploiting a Distributional Semantic Model for Disambiguating and Linking Entities in Tweets

Pierpaolo Basile
University of Bari Aldo Moro
pierpaolo.basile@uniba.it

Annalina Caputo
University of Bari Aldo Moro
annalina.caputo@uniba.it

Giovanni Semeraro
University of Bari Aldo Moro
giovanni.semeraro@uniba.it

Fedelucio Narducci
University of Bari Aldo Moro
fedelucio.narducci@uniba.it

ABSTRACT

This paper describes the participation of the UNIBA team in the Named Entity rEcognition and Linking (NEEL) Challenge. We propose a knowledge-based algorithm able to recognize and link named entities in English tweets. The approach combines the simple Lesk algorithm with information coming from both a distributional semantic model and usage frequency of Wikipedia concepts. The algorithm performs poorly in the entity recognition, while it achieves good results in the disambiguation step.

Keywords

Named Entity Linking, Distributional Semantic Models, Lesk Algorithm

1. INTRODUCTION

In this paper we describe our participation in the Named Entity rEcognition and Linking (NEEL) Challenge [4]. The task is composed of three steps: 1) identify entities in a tweet; 2) link entities to appropriate concepts[1] in DBpedia; 3) cluster entities that belong to specific classes (entity types) defined by the organizers.

We propose two approaches that share the same methodology to disambiguate entities, while differing in the approach used to recognize entities in the tweet. We implement two algorithms for entity detection. The former (*UNIBAsup*) exploits PoS-tag information to detect a list of candidate entities, while the latter (*UNIBAunsup*) tries to find sequences of tokens (n-grams) that are titles of Wikipedia pages or surface forms which refer to Wikipedia pages.

The disambiguation and linking steps rely on a knowledge-based method that combines a Distributional Semantic Models (DSM) with the prior probability assigned to each DBpedia concept. A DSM represents words as points in a mathematical space; words represented close in this space are similar. The word space is built analyzing word co-occurrences in a large corpus. Our algorithm is able to disambiguate an entity by computing the similarity between the context and the glosses associated with all possible entity concepts. Such similarity is computed through the vector similarity in the DSM. Section 2 provides details about the adopted strategies for: 1) Entity Recognition and 2) Linking. The experimental evaluation, along with commentary about results, are presented in Section 3.

[1] An entity can belong to several concepts.

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

2. THE METHODOLOGY

Our methodology is a two-step algorithm consisting in an initial identification of all possible entities mentioned in a tweet followed by the linking (disambiguation) of entities through the disambiguation algorithm. DBpedia is exploited twice in order to 1) extract all the possible surface forms related to entities, and 2) retrieve glosses used in the disambiguation process. In this case we use as gloss the extended abstract assigned to each DBpedia concept.

2.1 Entity Recognition

In order to speed up the entity recognition step we build an index where each surface form (entity) is paired with the set of all its possible DBpedia concepts. The index is built by exploiting Lucene API[2], specifically for each surface form (lexeme) occurring as the title of a DBpedia concept[3], a document composed of two fields is created. The first field stores the surface form, while the second one contains the list of all possible DBpedia concepts that refer to the surface form in the first field. The entity recognition module exploits this index in order to find entities in a tweet. Given a tweet, the module performs the following steps: 1) Tokenizing and PoS-tagging the tweet via Tweet NLP[4]; 2) Building a list of candidate entity. We exploit two approaches: all n-grams up to five words (*UNIBAunsup*); all sequences of tokens tagged as proper nouns by the PoS tagger (*UNIBAsup*); 3) Querying the index and retrieving the list of the top 25 matching surface forms for each candidate entity; 4) Scoring each surface form as the linear combination of: a) the score provided by the search engine; b) a string similar-

[2] http://lucene.apache.org/

[3] We extend the list of possible surface forms using also the resource available at: http://wifo5-04.informatik.uni-mannheim.de/downloads/datasets/

[4] http://www.ark.cs.cmu.edu/TweetNLP/

ity function based on the Levenshtein Distance between the candidate entity and the surface form in the index; c) the Jaccard Index in terms of common words between the candidate entity and the surface form in the index; 5) Filtering the candidate entities recognized in the previous steps: entities are removed if the score computed in the previous step is below a given threshold. In this scenario we set the threshold to 0.85. The output of the entity recognition module is a list of candidate entities in which a set of possible DBpedia concepts is assigned to each surface form in the list.

2.2 Linking

We exploit an adaptation of the distributional Lesk algorithm proposed by Basile et al. [1] for disambiguating named entities. The algorithm replaces the concept of word overlap initially introduced by Lesk [2] with the broader concept of semantic similarity computed in a distributional semantic space. Let $e_1, e_2, ...e_n$ be the sequence of entities extracted from the tweet, the algorithm disambiguates each target entity e_i by computing the semantic similarity between the glosses of concepts associated with the target entity and its context. This similarity is computed by representing in a DSM both the gloss and the context as the sum of words they are composed of; then this similarity takes into account the co-occurrence evidences previously collected through a corpus of documents. The corpus plays a key role since the richer it is the higher is the probability that each word is fully represented in all its contexts of use. We exploit the word2vec tool[5] [3] in order to build a DSM, by analyzing all the pages in the last English Wikipedia dump[6]. The correct concept for an entity is the one whose gloss maximizes the semantic similarity with the word/entity context. The algorithm consists of four steps.

1. Building the glosses. We retrieve the set $C_i = \{c_{i1}, c_{i2}, \ldots, c_{ik}\}$ of DBpedia concepts associated to the entity e_i. For each concept c_{ij}, the algorithm builds the gloss representation g_{ij} by retrieving the *extended abstract* from DBpedia.
2. Building the context. The context T for the entity e_i is represented by all the words that occur in the tweet except for the surface form of the entity.
3. Building the vector representations. The context T and each gloss g_{ij} are represented as vectors (using the vector sum) in the DSM.
4. Sense ranking. The algorithm computes the cosine similarity between the vector representation of each extended gloss g_{ij} and that of the context T. Then, the cosine similarity is linearly combined with a function that takes into account the usage of the DBpedia concepts. We analyse a function that computes the probability assigned to each DBpedia concept given a candidate entity. The probability of a concept c_{ij} is computed as the number of times the entity e_i is tagged with the concept c_{ij} in Wikipedia. Zero probabilities are avoided by introducing an additive (Laplace) smoothing.

We exploit the *rdf:type* relation in DBpedia to map each DBpedia concepts to the types defined in the task. In particular, we provide a manual map for all the types defined in the *dbpedia-owl* ontology to the respective types provided by the organizers.

[5]https://code.google.com/p/word2vec/

[6]We use 400 dimensions for vectors analysing only terms that occur at least 25 times.

3. EVALUATION AND RESULTS

This section reports results of our system on the development set provided by the organizers. The dataset consists of 500 manually annotated tweets. Results are reported in Table 1. The first column shows the entity recognition strategy, the other columns report respectively the F-measure of: strong link match (SLM), strong typed mention match (STMM), mention ceaf (MC). SLM measures the linking performance, while STMM takes into account both link and type. MC measures both recognition and classification.

ER Strategy	F-SLM	F-STMM	F-MC
$UNIBAsup$	0.362	0.267	0.389
$UNIBAunsup$	0.258	0.191	0.306

Table 1: Results on the development set

We cannot discuss the quality of the overall performance since we have not information about both baseline and other participants. However, we can observe that the recognition method based on PoS-tags obtains the best performance. We performed an additional evaluation in which we removed the entity recognition module and took entities directly from the gold standard. The idea is to evaluate only the linking step. Results of this evaluation are very encouraging, we obtain a F-SLM=0.563, while excluding the NIL instances we achieve a link match of 0.825. These results prove the effectiveness of the proposed disambiguation approach based on DSM.

Acknowledgments

This work fulfils the research objectives of the PON project EFFEDIL (PON 02_00323_2938699). The computational work has been executed on the IT resources made available by two PON projects financed by the MIUR: ReCaS (PONa3_00052) and PRISMA (PON04a2_A).

4. REFERENCES

[1] P. Basile, A. Caputo, and G. Semeraro. An Enhanced Lesk Word Sense Disambiguation Algorithm through a Distributional Semantic Model. In *Proc. of COLING 2014: Technical Papers*, pages 1591–1600. ACL, August 2014.

[2] M. Lesk. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proc. of SIGDOC '86*, SIGDOC '86, pages 24–26. ACM, 1986.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proc. of ICLR Work.*, 2013.

[4] G. Rizzo, A. E. Cano Basave, B. Pereira, and A. Varga. Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge. In M. Rowe, M. Stankovic, and A.-S. Dadzie, editors, *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 44–53, 2015.

AMRITA - CEN@NEEL : Identification and Linking of Twitter Entities

Barathi Ganesh H B, Abinaya N, Anand Kumar M, Vinayakumar R, Soman K P
Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Coimbatore, India
barathiganesh.hb@gmail.com, abi9106@gmail.com, m_anandkumar@cb.amrita.edu
vinayakumarr77@gmail.com, kp_soman@amrita.edu

ABSTRACT

A short text gets updated every now and then. With the global upswing of such micro posts, the need to retrieve information from them also seems to be incumbent. This work focuses on the knowledge extraction from the micro posts by having entity as evidence. Here the extracted entities are then linked to their relevant DBpedia source by featurization, Part Of Speech (POS) tagging, Named Entity Recognition (NER) and Word Sense Disambiguation (WSD). This short paper encompasses its contribution to #Micropost2015 - NEEL task by experimenting existing Machine Learning (ML) algorithms.

Keywords

CRF, Micro posts, NER

1. INTRODUCTION

Micro posts are a pool of knowledge with scope in business analytics, public consensus, opinion mining, sentimental analysis and author profiling and thus indispensable for Natural Language Processing (NLP) researchers. People use short forms and special symbols for easily conveying their message due to the limited size of micro posts which has eventually built complexity for traditional NLP tools [3]. Though there are number of tools, most of them rely on least ML algorithms which are effective for long texts than short texts. Thus by providing suffcient features to these algorithms the objective can be achieved. We experimented the NEEL task with the available NLP tools to evaluate their effect on entity recognition by providing special features available in tweets.

2. SELECTION OF ALGORITHMS

2.1 Tokenization

Tokenizing becomes highly challenging in micro posts due to the absence of lexical richness. It includes special symbols (:-), #, @user), abbreviations, short words (lol, omg), misspelled words, repeated punctuations and unstructured words (goooood nightttt, helloooo). Hence these micro posts were fed to the dedicated twitter tokenizer which accounts language identification, a lookup dictionary for list of names, spelling correction and special symbols [4][5] for effective tokenization.

2.2 POS Tagger

Due to the conversional nature of micro blogs with non-syntactic structure it becomes difficult in utilizing general algorithms with traditional POS tags in Penn Treebank and Wall Street Journal Corpus [6]. O'Conner et al. used 25 POS tagset which includes dedicated tags (@user, hash tag, G, URL, etc.) for twitter and reports 90% accuracy on POS tagging [7]. The ability of resolving independent assumptions and overcoming biasing problems make CRF as promised supervised algorithm for sequence labeling applications [8]. TwitIE tagger: which utilizes CRF to build the POS tagging model was thus used.

2.3 Named Entity Recognizer

CRF and SVM produced promising outcome for sequence labeling task which prompted us to use the same for our experiment. Long range dependency of the CRF can also solve Word Sense Disambiguation (WSD) problem over other graphical models by avoiding label and casual biasing during learning phase. Both CRF and SVM allow us to utilize the complicated feature without modeling any dependency between them. SVM is also well suited for sequence labeling task since learning can be enhanced by incorporating cost models [9]. These advantages provide flexibility in building expressive models with CRF suite and MALLET tools [10][11].

3. EXPERIMENTS AND OBSERVATION

The experiment is conducted on i7 processor with 8GB RAM and the flow of experiment is shown in Figure 1. The training dataset consists of 3498 tweets with the unique tweet id. These tweets have 4016 entities with 7 unique tags namely Character, Event, Location, Organization, Person, Product and Thing [1][2]. POS tag for the NER is obtained from TwitIE tagger after tokenization which takes care of the nature of micro posts and provides an outcome desired by the POS tagger model. The tags are mapped to BIO Tagging of named entities. Considering the entity as a phrase, token at the beginning of the phrase is tagged as 'B-(original tag)' and the token inside the phrase is tagged as 'I-(original tag)'. Feature vector constructed with POS tag and addi-

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

tional 34 features like root word, word shapes, prefix and suffix of length 1 to 4, length of the token, start and end of the sentence, binary features - whether the word contains uppercase, lower case, special symbols, punctuations, first letter capitalization, combination of alphabet with digits, punctuations and symbols, token of length 2 and 4 , etc.
After constructing the feature vector for individual tokens in the training set and by keeping bi-directional window of size 5, the nearby token's feature statistics are also observed to help the WSD. The final windowed training sets are passed to the CRF and SVM algorithms to produce the NER model. The development data has 500 tweets along with their id and 790 entities [1][2]. The development data is also tokenized, tagged and feature extracted as the training data for testing and tuning the model. The developed model performance

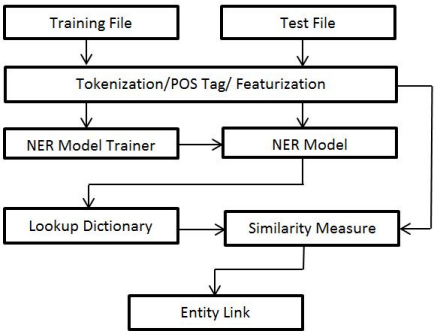


Figure 1: Overall Model Structure

is evaluated by 10- fold cross validation of training set and validated against the development data. The accuracy is computed as ratio of total number of correctly identified entities to the total number of entities and tabulated in Table 1.

$$Accuracy = \frac{\sum correctly\ identified\ entities}{total\ entities} \times 100 \quad (1)$$

MALLET incorporates O-LBFGS which is well suited for log-linear models but shows reduced performance when compared to CRFsuite which engulfs LBFGS for optimization [12][13]. SVM's low performance can be improved by increasing the number of features which will not introduce any over fitting and sparse matrix problem [9].
The final entity linking part is done by utilizing lookup dictionary (DBpedia 2014) and sentence similarity. The entity's tokens are given to the look up dictionary which results in few related links. The final link assigned to the entity is based on maximum similarity score between related links and proper nouns in the test tweet. Similarity score is computed by performing dot product between unigram vectors of proper nouns in the test tweet and the unigram vectors of related links from lookup dictionary. Entity without related links is assigned as NIL.

4. DISCUSSION

This experimentation is about sequence labeling for entity identification from micro posts and extended with DBpedia resource linking. By observing Table 1, it is clear that CRF shows great performance and paves way for building a smart NER model for streaming data application. Even though CRF seems to be reliable, it is dependent on the feature

Table 1: Observations

Tools	10 Fold-Cross Validation	Development Data	Time (mins)
Mallet	84.9	82.4	168.31
SVM	79.8	76.3	20.15
CRFSuite	88.9	85.2	4.12

that has direct relation with NER accuracy. The utilized TwitIE tagger shows promising performance in both the tokenization and POS tagging phases. The special 34 features extracted from the tweets improves efficacy by nearing 13% greater than the model with absence of special features. At linking part, this work is limited using dot product similarity which could be improved by including semantic similarity.

5. REFERENCES

[1] Rizzo, Giuseppe and Cano Basave, Amparo Elizabeth and Pereira, Bianca and Varga, Andrea, *Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge.*, In 5th Workshop on Making Sense of Microposts (#Microposts2015), pp. 44–53, 2015.

[2] Matthew Rowe and Milan Stankovic and Aba-Sah Dadzie, *Proceedings, 5th Workshop on Making Sense of Microposts (#Microposts2015): Big things come in small packages, Florence, Italy, 18th of May 2015*, 2015.

[3] Dlugolinsky S, Marek Ciglan and M Laclavik, *Evaluation of named entity recognition tools on microposts*, INES, 2013 , pp. 197-202. IEEE, 2013.

[4] Bontcheva K, Derczynski L, Funk A, Greenwood M A, Maynard D, and Aswani N, *TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text*, In RANLP, pp. 83-90, 2013, September.

[5] Brendan O'Connor, Michel Krieger and David Ahn, *TweetMotif: Exploratory Search and Topic Summarization for Twitter*, ICWSM, 2010.

[6] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller and Justin Martineau, *Annotating named entities in Twitter data with crowdsourcing*, 2010.

[7] Kevin Gimpel, et al, *Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments*, HLT'11, 2011.

[8] John Lafferty,Andrew McCallum and Fernando Pereira, *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*, 2001.

[9] Chun-Nam John Yu, Thorsten Joachims, Ron Elber and Jaroslaw Pillardy, *Support vector training of protein alignment models*, in Research in Computational Molecular Biology, 2007.

[10] Naoaki Okazaki, *CRFsuite: a fast implementation of Conditional Random Fields (CRFs)*, 2007.

[11] McCallum and Andrew Kachites, *MALLET: A Machine Learning for Language Toolkit*, http://mallet.cs.umass.edu, 2002.

[12] Galen Andrew and Jianfeng Gao, *Scalable Training of L1-Regularized Log-Linear Models*, ICML, 2007.

[13] Jorge Nocedal, *Updating Quasi-Newton Matrices with Limited Storage*, Mathematics of Computation, Volume 35, Number151, pp:773-782, 1980.

Named Entity Extraction and Linking in #Microposts

Priyanka Sinha
TCS Innovation Lab Kolkata
Indian Institute of Technology Kharagpur
priyanka27.s@tcs.com

Biswanath Barik
TCS Innovation Lab Kolkata
Tata Consultancy Services Limited
biswanath.barik@tcs.com

ABSTRACT
The task of Named Entity Extraction and Linking (NEEL) challange 2015 [5] is considered as two successive tasks : Named Entity Extraction (NEE) from the tweets and Named Entity Linking (NEL) with DBpedia. For NEE task we use CRF++ [1] to create a language model on the given training data. For entity linking, we use DBpedia Spotlight.

Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: Miscellaneous

General Terms
Experiment

Keywords
Twitter, Entity, Linking, Social Media, DBpedia

1. INTRODUCTION
Information Extraction (IE) from short messages or microblogs like tweets is an emerging field of research due to its commercial applications like ecommerce, recommendation etc. and social administration like social security. Entity linking (or entity resolution) is one such task which deals with identifying and extracting the Named Entities that belong to the tweets and disambiguating them by linking to the correct reference entities in the knowledge base.

The entity linking problem is well explored on normal text. However, the existing techniques of entity linking do not work well on short messages as the microblogs do not have sufficient context to classify (or disambiguate) the mentions. In this work we have identified the mention by creating an entity recognition model on the given training data and link them to the DBpedia using DBpedia Spotlight.

The rest of the paper is organized as follows: Section 2 describes our proposed approach which includes data preparation and feature selection for named entity recognition model creation and entity linking method. Section 3 describes the setup for web access. The result of our work is discussed in Section 4. Section 5 illustrates the future scope of our work followed by the references.

Published as part of the #Microposts2015 Workshop proceedings, available online as CEUR Vol-1395 (http://ceur-ws.org/Vol-1395)

#Microposts2015, May 18th, 2015, Florence, Italy.

2. METHODOLOGY
In our approach we have divided the Named Entity Extraction and Linking (NEEL) [5] task into two consecutive subtasks, namely, Named Entity Extraction and Named Entity Linking.

2.1 Named Entity Extraction
The NER task is viewed here as a sequence labeling problem. Given an input tweet, this step aims to identify the word sequences that constitute a Named Entity and classify each such entity into one of the predefined classes. For entity recognition and classification task, we have developed a model on the given training data using Conditional Random Fields (CRFs) which is an undirected graphical model used mainly for sequence labeling.

As we have discussed in the previous section that the context of the tweets is short, sometimes noisy and informal and thus, their syntactic structures are not always comparable to the normal texts. [4] showed that the Part-of-Speech (POS) features of surface tokens, Shallow Parsing (or Chunking) information, Capitalization indicators etc. are useful for improving NE recognition from tweets provided these modules should be trained on twitter data. In this experiment, we have added POS tag information to the training data using Twitter NER[3], used word features and some binary features like punctuations, digits, dots, hashtags, @, capitalization indicators, existence of URLs, underscore, hyphen etc. as features indicating or not indicating NEs for training NE recognition model. We were motivated to use [3] as it allows to tokenize and distinguish between nouns and other punctuations and tweet related artefacts well. We used [1] as it was relatively simple to adapt to our task.

2.1.1 Data Preparation
In the data preparation step, we have identified the word sequences refering to a Named Entity(NE) in the training data using the gold standard. The training data is tokenized, part-of-speech(POS) tagged using Twitter NER[3] and converted into 'BIO' format. For example, the NEs identified in the tweet ID: 100678378755067904, tweet "RT @HadleyFreeman: NOTHING on US news networks about

London riots. Can you imagine the BBC ignoring, say, riots in NYC? #americanewsfail" as follows

```
RT ~ O
@HadleyFreeman @ B-Person
: ~ O
NOTHING N O
on P O
US ^ B-Location
news N O
networks N O
about P O
London ^ B-Event
riots N I-Event
. , O
Can V O
you O O
imagine V O
the D O
BBC ^ B-Organization
ignoring V O
, , O
say V O
, , O
riots N O
in P O
NYC ^ B-Location
? , O
#americanewsfail # O
```

2.1.2 *Feature Selection*

We have experimented with various feature types, various window lengths and their combinations and come up with the following feature set which gave us a good result. We experimented with some context window lengths and 5 gave us good results.

- Contextual (Word) Features: a context window of size five: W_{i-2} W_{i-1} W_i W_{i+1} W_{i+2}
- Part-of-Speech (POS) Features: a context of size five: P_{i-2} P_{i-1} P_i P_{i+1} P_{i+2}
- Word having Capitalization: binary feature
- Word having Punctuation: binary feature
- Is a Digit: binary feature
- Word having a Dot: binary feature
- Word having hashtag: binary feature
- Word having @: binary feature

2.2 Named Entity Linking

For linking, we use the annotations returned by DBpedia Spotlight REST API as the candidates and look for the longest matching surface forms.

We take the output of the NEE task and collect the named entities that are extracted and their categories. To identify correct start position we check for # and @. For each tweet, using the B/I tags we find the longest consecutive entities that make up a single entity. For example, in the tweet above, "London riots" would be treated as a single entity. For each tweet, DBpedia Spotlight REST API is accessed with confidence and support set to 0 with accepted return text in XML. We use the DBpedia Spotlight's annotate endpoint to obtain all the links at once. For each entity returned from DBpedia Spotlight, if the surface form is found to be a substring of any of the entities and if a substring match is found the corresponding URI is returned. For named entities for which no match is found, if it is an existing nil entity then the nil id is returned, else the nil counter is incremented and returned.

3. SETUP

We used perl for transforming the data. We used the CMU Twitter NLP[3] package for generating POS, CRF++[1] package and DBpedia Spotlight[2] REST API.

3.1 Web access

We use JSP to create our REST API, which uses perl which in turn uses curl to connect to DBpedia Spotlight[2] REST endpoints.

4. EVALUATION

The precision for strong link match with the training set itself is 30.49%, recall is 30.29% and f1 is 30.39%. For the tagging of correct entity type the precision with the training set itself is 82.89%, recall 82.35% and f1 82.62%.

The precision for strong link match with the development set is 14.82%, recall is 7.97% and f1 is 10.37%. For the tagging of correct entity type the precision with the training set itself is 41.65%, recall 22.41% and f1 29.14%.

5. FUTURE WORK

As we can see using the CMU POS tagger[3] and CRF[1] discovers the entities well, but the way we do linking needs more work.

6. REFERENCES

[1] Crf++: Yet another crf toolkit.

[2] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*, 2013.

[3] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *In Proceedings of NAACL*, 2013.

[4] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, Scotland, UK, July 2011.

[5] G. Rizzo, A. E. Cano Basave, B. Pereira, and A. Varga. Making Sense of Microposts (#Microposts2015) Named Entity rEcognition and Linking (NEEL) Challenge. In M. Rowe, M. Stankovic, and A.-S. Dadzie, editors, *5th Workshop on Making Sense of Microposts (#Microposts2015)*, pages 44–53, 2015.