Using Arabic Twitter to support analysis of the spread of Infectious Diseases

Lama Saleh Alsudias, BSc (Hons), MSc
School of Computing and Communications
Lancaster University

A thesis submitted for the degree of
Doctor of Philosophy

May, 2022

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. This thesis does not exceed the maximum permitted word length of 80,000 words including appendices and footnotes, but excluding the bibliography. A rough estimate of the word count is: 29319

Lama Saleh Alsudias

Using Arabic Twitter to support analysis of the spread of Infectious Diseases

Lama Saleh Alsudias, BSc (Hons), MSc.
School of Computing and Communications, Lancaster University

A thesis submitted for the degree of *Doctor of Philosophy.* May, 2022

Abstract

This study investigates how to use Arabic social media content, especially Twitter, to measure the incidence of infectious diseases. People use social media applications such as Twitter to find news related to diseases and/or express their opinions and feelings about them. As a result, a vast amount of information could be exploited by NLP researchers for a myriad of analyses despite the informal nature of social media writing style. Systematic monitoring of social media posts (infodemiology or infoveillance) could be useful to detect misinformation outbreaks as well as to reduce reporting lag time and to provide an independent complementary source of data compared with traditional surveillance approaches. However, there has been a lack of research about analysing Arabic tweets for health surveillance purposes, due to the lack of Arabic social media datasets in comparison with what is available for English and some other languages. Therefore, it is necessary for us to create our own corpus. In addition, building ontologies is a crucial part of the semantic web endeavour. In recent years, research interest has grown rapidly in supporting languages such as Arabic in NLP in general but there has been very little research on medical ontologies for Arabic.

In this thesis, the first and the largest Arabic Twitter dataset in the area of health surveillance was created to use in training and testing in the research studies presented. The Machine Learning algorithms with NLP techniques especially for Arabic were used to classify tweets into five categories: academic, media, government, health professional, and the public, to assist in reliability and trust judgements by taking into account the source of the information alongside the content of tweets. An Arabic Infectious Diseases Ontology was presented and evaluated as part of a new method to bridge between formal and informal descriptions of Infectious Diseases. Different qualitative and quantitative studies were performed to analyse Arabic tweets that have been written during the pandemic, i.e. COVID-19, to show how Public Health Organisations can learn from social media. A system was presented that measures the spread of two infectious diseases based on our Ontology to illustrate what quantitative patterns and qualitative themes can be extracted.

Using Arabic Twitter to support analysis of the spread of Infectious Diseases

Lama Saleh Alsudias, BSc (Hons), MSc.
School of Computing and Communications, Lancaster University
A thesis submitted for the degree of *Doctor of Philosophy*. May, 2022

Abstract in Arabic

تبحث هذه الدرَاسة عن كيفية استخدَام محتوَى وسَائل التوَاصل الَاجتمَاعي العربية، وخَاصة تويتر لقيَاس الَاصَابة بَالَامرَاض المعدية. يستخدم الَاشخَاص تطبيقَات الوسَائط الَاجتمَاعية مثل تويتر للّعثور علَى الَاخبَار المتعلقة بَالَامرَاض و/او للّتعبير عن آرَاءهم و مشَاعرهم عنهَا. نتيجة لذلك، يمكن للّبَاحثين استغلَال هذَا القدر الهَائل من المعلومَات لَاجرَاء عد د لَا يحصَى من التحليلَات علَى الرغم من الطريقة غير الرسمية لَاسلوب الكتَابة علَى وسَائل التوَاصل الَاجتمَاعي. من الـمـمكن ان يكون الرصد المنتظم لمنشورَات وسَائل التوَاصل الَاجتمَاعي (المرَاقبة) مفيدَا للّكشف عن تفشي المعلومَات الخَاطِئة و كذلك لتقليل التَاخر في الَابلَاغ عن الَاصَابَات و توفير مصدر مستقل لـلبيَانَات مقَارنة بَاسَاليب المرَاقبة التقليدية. و مع ذلك، يوجد نقص في الَابحَاث حول تحليل التغريدَات العربية لَاغرَاض المرَاقبة الصحية، بسبب نقص بيَانَات وسَائل التوَاصل الَاجتمَاعي العربية مقَارنة بمَا هو متوفر بَالـلغة الَانجليزية و بعض الـلغَات الَاخرَى. لذلك، من المهم بَالنسبة لنَا، انشَاء مجموعة بيَانَات خَاصة بنَا. ايضَا يعد بنَاء الَانطولوجيَا جزآ مهمَا من مسعَى الويب الدلَالي. في السنوَات الَاخيرة، نمَا الَاهتمَام البحثي بسرعة في دعم الـلغَات مثل العربية في معَالجة الـلغة الطبيعية بشكل عَام ولكن القليل منهَا حول الَانطولوجيَا الطبية لـلغة العربية.

في هذه الرسَالة، تم انشَاء اول و اكبر مجموعة بيَانَات عربية علَى تويتر في مجَال المرَاقبة الصحية لَاستخدَامهَا في تدريب و اختبَار النظَام. تم استخدَام خوَارزميَات التعلم الَالي مع تقنيَات معَالجة الـلغة الطبيعية الخَاصة بَالـلغة العربية لتصنيف التغريدَات الَى خمس فِئَات: الَاكَاديمية، وَالَاعلَامية، و الحكومية، و الصحية المتخصـصة، و العَامة، لـلمسَاعدة في الحكم علَى ثقة التغريدة من خلَال مصدر التغريدة الَى جَانب محتوَاهَا. ايضَا تم بنَاء و تقيـيم انطولوجيَا لـلَامرَاض المعدية بَالـلغة العربية كجزء من طريقة جديدة لـلربط بين الَاوصَاف الرسمية

و الغير رسمية لـلَامرَاض المعدية. تم اجرَاء عدة درَاسَات نوعية و كمية لتحليل التغريدَات العربية التي تمت كتَابتهَا اثنَاء الوبَاء، كوفيد٩١، لَاظهَار كيف يمكن لمنظمَات الصحة العَامة التعلم من وسَائِل التوَاصل الَاجتمَاعي. تم تقديم نظَام يقيس انتشَار مرضيين معديين بنَاء علَى الَانطولوجيَا لتوضيح الَانمَاط الكمية و الموضوعَات النوعية التي يمكن استخلَاصهَا.

Publications

The following publications have been published in major conferences and a journal while developing this thesis, and to an extent has guided the thesis into what it has become:

- Lama Alsudias and Paul Rayson (2019). "Classifying information sources in Arabic twitter to support online monitoring of infectious diseases". In: *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pp. 22–30

- Lama Alsudias and Paul Rayson (May 2020c). "Developing an Arabic Infectious Disease Ontology to Include Non-Standard Terminology". In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4844–4852. URL: `https://www.aclweb.org/anthology/2020.lrec-1.596`

- Lama Alsudias and Paul Rayson (July 2020b). "COVID-19 and Arabic Twitter: How can Arab World Governments and Public Health Organizations Learn from Social Media?" In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics. URL: `https://www.aclweb.org/anthology/2020.nlpcovid19-acl.16`

- Lama Alsudias and Paul Rayson (Sept. 2021b). "Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study". In: *JMIR Med Inform* 9.9, e27670. ISSN: 2291-9694. DOI: `10.2196/27670`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/34346892`

In addition, I presented these posters while working on this thesis:

- Lama Alsudias and Paul Rayson (Apr. 2020a). "Analyzing the Spread of Influenza in Arabic Twitter". English. In: HEALTHCARE TEXT ANALYTICS CONFERENCE; Conference date: 23-04-2020. URL: `http://healtex.org/healtac-2020/`

- Lama Alsudias and Paul Rayson (June 2021a). "Detecting COVID-19 Misinformation in Arabic Tweets". English. In: HEALTHCARE TEXT ANALYTICS CONFERENCE 2021; Conference date: 17-06-2021 Through 18-06-2021. URL: `http://healtex.org/healtac-2021/`

Acknowledgements

First and foremost, I must thank Allah (Almighty God) for providing me with the strength, knowledge, and fortitude to persevere through all of the challenges that I encountered during my studies, and for blessing me with the ability to complete my job successfully.

This thesis and all of my triumphs are dedicated to my parents, Mr. Saleh and Mrs. Aishah, who have supported me spiritually and physically throughout my life; without their encouragement and support, this work and all of my successes would not have been completed. I would like to express my gratitude to them for their love, prayers, support, and faith in me to complete my study abroad program.

Professor Paul Rayson, my supervisor, deserves special thanks for assisting me with ideas and for his ongoing support and admiration for all of the work I have given him. Thanks to his encouragement to publish, I have seen several of my findings have been mentioned in other studies. He is an example of a genuine excellent man since he appreciated all of my situations.

Special thanks to my husband, Adel, who always supports me in every stage of my life. I know that this work would not exist without his presence in my life and his sacrifices.

I would also want to thank my sister and brothers for their faith in me, as well as their encouragement and support. They have been such an inspiration to me, and I would like to express my gratitude to everyone of them.

Last but not least, I would like to thank my lovely daughters, Hoor, Farah, and Raghad, for their smiles or even tears that encouraged me to finish this work, for sharing with me the ups and downs of alienation; they sacrificed even though they are young to understand that.

Thanks go to King Saud University, Riyadh, Saudi Arabia for giving me a scholarship to study in the UK.

Finally, I would want to express my gratitude to my relatives, friends, and everyone at the School of Computing and Communications for their cooperation and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since the advent of Web 2.0, most people participate in the Web to express their opinions, which also gives researchers the opportunity to analyse those opinions across large scale populations more than is otherwise possible. One aim of this process is to summarise general opinions regarding international, national, or local events or themes in the huge amount of data available on the Internet. Microblogging platforms are also used by people and organisations to share information on many different topics, and these can also be a source of emergency and/or vital public health information. Due to social media's popularity as a communication platform for real time messages, it could potentially be useful to analyse tweets for health surveillance purposes. We hypothesise that this may reduce reporting lag time compared to the traditional approaches for monitoring infectious diseases, while also providing qualitative and quantitative information of how people talk about infectious diseases. The need to detect the spread of infectious diseases as soon as possible via Arabic Twitter also provides motivation for this PhD thesis.

1.1 Background

There are three key aspects underpinning this research which we briefly introduce here: Arabic language, infectious diseases, and Twitter.

The Arabic language is the world's fifth most widely spoken language. It has several characteristics that make it more difficult to analyse than other languages, and Natural Language Processing (NLP) tools and methods are not as well established as they are for English. Frequently encountered issues include words with multiple meanings and Arabic dialect identification in non-standard contexts. To clarify, formal medical terminology uses modern standard Arabic while informal language is used in the social media which leads to ambiguity in understanding a certain terminology. More details about the Arabic language and its challenges can be found in Section 2.2. Moreover, Section 7.1 contains further examples of the discrepancy in terminology between casual language on social media and official medical terminology.

Infectious diseases are illnesses that may be spread from one organism to another by the virus or germ that causes them. While epidemics and pandemics have historically had significant social and economic consequences for afflicted communities, in today's linked globe, the consequences are really global. This point was made abundantly obvious by the COVID-19 pandemic that began in 2020. Infections can readily move from one place to another. Until the virus is completely eradicated from the world, a rise in cases in one region might result in a rebound of instances in other parts of the world. In addition, COVID-19 became the leading cause of death in 2020 [1]. Section 2.1 has further information about the infectious diseases and epidemics and pandemics. It also discusses some NLP past studies that focus on other infectious diseases.

In Twitter, users post and engage with messages known as "tweets", which are microblogging and social networking services combined. Tweets can be posted, liked, and retweeted by users. Thirty four distinct languages are supported by Twitter for microblogging[2]. In addition, there are more opportunities for bending or breaking more formal grammar rules or standard spelling in writing on social media. In other words, users tend to use more informal language when they tweet. This has led to many techniques needing to be used by academic researchers when they try to analyse tweets as a result of using Twitter as data source as explained in Section 2.3.1. For a very wide range of topics, academic researchers have analysed tweets using the facilities provided by Twitter's developer tools. Importantly, when carrying out such research there are some restrictions and rules when collecting tweets related to the platform's terms and conditions as discussed in Section 3.2 and ethical considerations, see section 3.3.

1.2 Motivation and Objectives

Monitoring the spread of infectious diseases is important using a range of complementary approaches, formal and informal, due to their effect on human life. Formal strategies such as hospital records, laboratory statistics, or household surveys need time to produce results. On the other hand, informal methods like search engines or

[1] https://www.bcm.edu/departments/molecular-virology-and-microbiology/emerging-infections-and-biodefense/introduction-to-infectious-diseases

[2] https://developer.twitter.com/en/docs/twitter-for-websites/supported-languages

social media may detect queries or discussion of disease names near to real time. Governments now make extensive efforts to detect infectious diseases as soon as possible in order to provide medicines and prevent the spread of the disease. So, there is need for a system that enhances the quicker detection of diseases in the future. The results of this study, once completed, will be of interest to the World Health Organisation, Ministry of Health in Saudi Arabia [3], and Saudi Food and Drug Authority [4].

The main motivation of this work is to combine these challenges and investigate how to use Arabic social media content, especially Twitter, to measure the incidence of infectious diseases.

The research objectives are outlined as follows to guide the research and achieve the aims:

1. Collect data for analysis of the spread of infectious diseases.
2. Create and evaluate a classification system to identify information sources.
3. Create and evaluate an Arabic Infectious Diseases Ontology to be a bridge between formal and informal descriptions of infectious diseases.
4. Apply different quantitative and qualitative studies to analyse tweets during the COVID-19 pandemic.
5. Create a system that measures the spread of infectious diseases based on the new Ontology using state of the art NLP techniques like text classification, annotation, named entity recognition.
6. Create a system which can predict the approximate location of the infected person via a tweet's content, when no geo-location is assigned to the tweet.

1.3 Research Questions and Contributions

The overall questions being answered in this PhD is how can we improve the analysis of the spread of infectious diseases in Arabic speaking countries, with a focus on Saudi Arabia via Twitter using NLP methods. This overall question breaks down into the following research questions:

[3] https://www.moh.gov.sa

[4] https://www.sfda.gov.sa

1. How can we differentiate between the different sources of information on social media in the area of health surveillance?
2. In what way does the Infectious Diseases Ontology affect the collection and analysis of the spread of infectious disease?
3. What are the topics that are discussed during pandemics?
4. How can we measure the spread of infectious diseases in Arabic speaking countries, with a focus on Saudi Arabia via Twitter?
5. How is it possible to locate the location of the infected people using content of the tweet?

Here we will list the contributions made in this research and then go over them in depth later:

1. The creation of a new Arabic infectious diseases dataset.
2. The creation of a new Arabic Infectious Diseases Ontology.
3. The creation and evaluation of a classification system to find the source of information.
4. A study of quantitative patterns and qualitative themes of how people talk about infectious diseases on Arabic social media.
5. An analytical investigation of the spread of infectious diseases is conducted using informal language.
6. Using Arabic Named Entity Recognition (ANER) methods to predict the location of the user depending on the content of the tweet.

1.4 Structure of Thesis

This thesis is split into eight chapters as shown in the following:

- Chapter 1: **Introduction**
  provides background information about Arabic language, infectious diseases, and Twitter, the objectives of this research, and the contributions.

- Chapter 2: **Literature Review**
  defines the thesis's main work areas which are infectious diseases, Arabic Natural Language Processing (ANLP), analysing social media data, and Machine Learning (ML) algorithms. It also covers the current and past work within the area of analysing health-related Arabic and non-Arabic Twitter data and identifying geo-location from tweets.

- Chapter 3: **Data Collection**
  describes the data collection procedure, including Twitter rules, ethical approval, methodology, some statistics, and discussion of possible future possibilities within the dataset.

- Chapter 4: **Information Sources of Arabic Infectious Diseases Tweets**
  presents our methods to classify information sources in Arabic Twitter to support online monitoring of infectious diseases. This chapter answers the first research question (RQ1).

- Chapter 5: **An Arabic Infectious Disease Ontology**
  describes the Arabic Infectious Disease Ontology that was built to help with numerous essential applications, such as tracking the transmission of infectious diseases on social media. This chapter provided a solution to the second research question (RQ2).

- Chapter 6: **Analysing Tweets During a Pandemic**
  presents how Public Health Organisations (PHOs) can learn from social media data by analysing tweets during pandemic. It includes multiple qualitative and quantitative studies: analysing the subjects spoken between individuals at the peak of COVID-19, detecting COVID-19-related rumours, and predicting the source of COVID-19-related tweets. Research Question RQ3 was addressed in this chapter.

- Chapter 7: **Monitoring Infectious Diseases**
  introduces a new approach to analyse Arabic tweets to estimate their usefulness for health surveillance and understand the impact of the informal terms in the analysis. It also shows how to identify the location of the infected people using content of the tweet. Research questions RQ4 and RQ5 were answered in this chapter.

- Chapter 8: **Conclusion**
  summarizes the thesis achievements, conclusion and future work.

Chapter 2

Literature Review

This chapter presents a review of the key areas of work related to this thesis, namely: infectious diseases, Arabic natural language processing, analysing social media data, machine learning algorithms, and some previous works related to this thesis in parts of analysing health-related Arabic and non-Arabic Twitter data and identifying geo-location from tweets.

2.1 Infectious Diseases

2.1.1 Overview of Infectious Diseases

Infectious diseases are illnesses where the virus or microbe causing the disease can be transmitted to another organism of the same species (Litin and Nanda, 2009). Any infectious diseases are transmissible from one human to another. Insects and other animals may spread certain diseases. Others may be contracted by eating tainted food or drinking tainted water, or by being exposed to microbes in the atmosphere. Fever and exhaustion are typical signs and symptoms, which differ depending on the organism that is causing the infection. Mild infections may be treated with rest and home remedies, while more severe infections may necessitate hospitalisation. In the worst case, infectious diseases may cause death. Vaccines can eliminate certain contagious diseases, such as measles and chickenpox. Hand-washing regularly and properly also serves to shield you from the bulk of infectious diseases (Litin and Nanda, 2009).

The World Health Organization[1] statistics shown in Table 2.1 indicate the number of cases and deaths that are reported from countries in the Eastern Mediterranean

[1] http://www.emro.who.int

Region in 2017[2]. We can see that 1,019,044 people were infected by Cholera in Yemen as well as 2,237 death cases as an example for only one disease.

Disease	Country	No. of Cases	No. of Deaths
acute watery diarrhea	Sudan	36,460	818
Avian influenza (H5N1)	Egypt	359	122
Chickenpox (Varicella)	Pakistan	14,246	21
Chikungunya	Pakistan	8375	0
	Somalia	78,784	1159
Cholera	Yemen	1, 019,044	2237
	Iraq	505	3
Dengue	Pakistan	25,872	69
	Sudan	179	3
Diphtheria	Yemen	429	42
	Oman	2	0
Middle East respiratory syndrome (MERS)	Saudi Arabia	234	65
	United Arab Emirates	7	0

Table 2.1: Emerging infectious disease outbreaks reported from the countries of the Region in 2017

2.1.2 Epidemics and Pandemics

A disease that affects a large number of people in a community, population, or region is known as an epidemic. While, when the infectious disease spreads over different countries it becomes known as a pandemic. For instance, COVID-19 was an epidemic when confined to Wuhan, China but became a pandemic when it spread further into the world. For both types, people expect to be protected from any given disease and find accurate information about it.

[2] http://www.emro.who.int/pandemic-epidemic-diseases/news/emerging-infectious-disease-outbreaks-reported-in-the-eastern-mediterranean-region-in-2017.html

Influenza, which is commonly known flu, is one of the infectious diseases that caused by influenza viruses. For the majority of cases, the flu goes away on its own. However, flu and its symptoms can also be fatal. Although the influenza vaccine is not 100% successful every year, it is also the best protection against influenza. In an average year, 5-15% of the population worldwide get flu. Per year, there are 3-5 million serious cases, and up to 650,000 deaths worldwide. High-risk populations, including young children, seniors and individuals with serious illnesses most often are more likely to result in death. The influenza peaks in winter, while in the tropical areas influenza will occur all year round in temperate parts of the world. Every 10 to 50 years since the late 1800s there have been major outbreaks of new influenza strains termed pandemics. Since 1900 there are five flu pandemics: the Spanish influenza pandemic 1918-1920; Asian Flu in 1957; Hong Kong flu in 1968; Russian flu in 1977; and the swine flu pandemic in 2009 (*Influenza* 2021; Saunders-Hastings and Krewski, 2016).

In March 2020, the World Health Organization (WHO) announced COVID-19, an infectious disease caused by a coronavirus, outbreak as a pandemic. It can be transmitted through airborne respiration or nasal fluid droplets when an infected individual coughs or sneezes. When the virus appeared there was no treatment or vaccine and it caused a large number of deaths of people around the world. According to WHO, in March 2021, the confirmed cases reached 119,220,681 and the deaths were 2,642,826 around the world and in Saudi Arabia, there have been up to 382,407 cases recorded and 6,500 deaths on the same month. Figure 2.1 shows the number of cases by the geographical locations around the world. As a result, many governments during the pandemic took various decisions to protect citizens from COVID-19 such as online education, lockdown, quarantine, travel bans, social distancing, and self isolation. At the start of 2021, various vaccines were deployed in several countries and governments started encouraging people to take vaccines due to false information that spread via social media. This pandemic differs from those previously since people around the world shared very large quantities of information via social media about it. This led to huge amounts of data appearing online and gave researchers the opportunity to analyse them in multiple directions as explain in the next section (Section 2.1.3).

2.1.3 NLP and Infectious Diseases

Institutions and people with different specialities, including governments, health organisations, researchers and academics, commercial organisations, universities and schools, need to collaborate with each other to reduce the disease spread. Computer science researchers in a wide range of sub-fields have contributed useful studies and

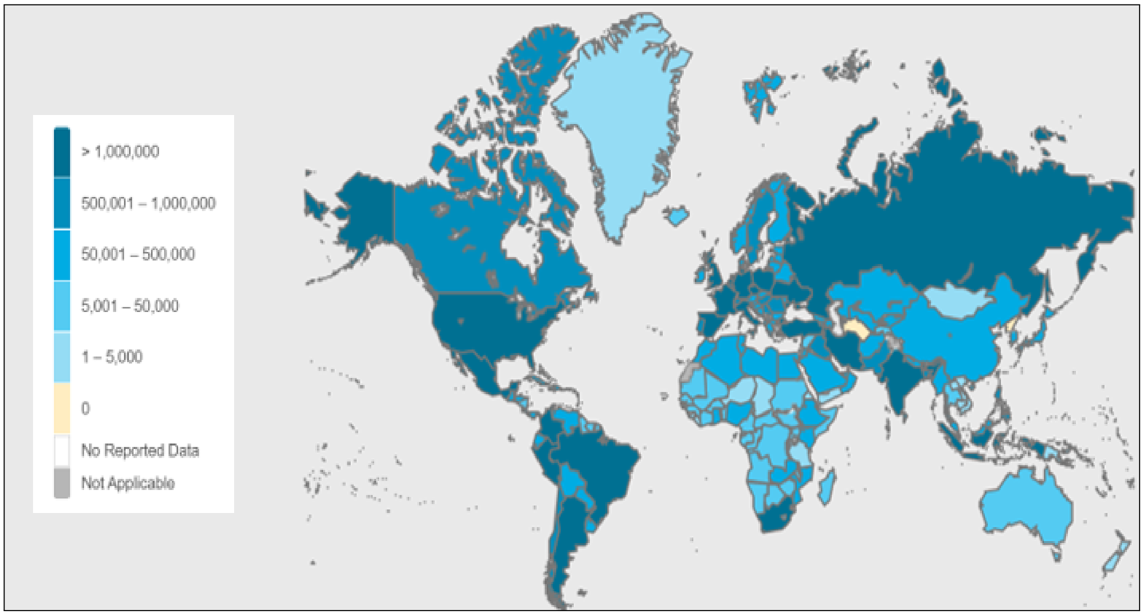


Figure 2.1: COVID-19 cases by geographical location. Adapted from WHO (`https://covid19.who.int/`)

applications that cross over with health concerns. For example, health informatics is a broad area of research that encompasses the design, creation, and implementation of technologies for the purpose of improving health care. It is one of the fastest growing sectors of the health care industry. Health informatics research focuses on the application of artificial intelligence to healthcare and the architecture of embedded medical equipment. Additionally, medical informatics encompasses contemporary developments of neuroinformatics and cognitive informatics in the areas of brain imaging and emulation. In some cases, the term informatics is often used to refer to the application of library science to hospital data processing (Nadri et al., 2017).

Biomedical text mining as an area of study combines concepts from natural language processing, bioinformatics, medical informatics, and computational linguistics. The techniques established in this field are frequently extended to the biomedical and molecular biology literature that is freely accessible through online resources e.g. MEDLINE and PubMed. For instance, NaCTem[3], The National Centre for Text Mining, is is a text mining centre in the UK, publicly supported. It works for many years and the long running BioNLP workshops from the ACL SIGBIOMED special interest group. Another example is the HealTAC[4], the UK healthcare text

[3] `http://www.nactem.ac.uk/`

[4] `http://healtex.org/`

analytics network, that extends research to other areas beyond biomedical. It is a multidisciplinary research group whose goal is to make it easier to use healthcare free-text, such as clinical notes, emails, social media posts, and literature, in research and clinical practise. Every year, it hosts a conference with keynote lectures, scientific papers, discussion panels, a business forum, tech demos, a PhD forum, and poster sessions. For example, the paper by Guidère (2020) uses the NLP in a suicide monitoring framework to apply machine learning and inspect social media. In particular in NLP, over recent years, there have been several research enterprises successfully studying various other epidemics such as listeria, influenza, swine flu, measles, meningitis, Ebola, chickenpox, and H1N1 (Ahmed, Peter Bath, et al., 2018; Aramaki, Maskawa, and Morita, 2011; Jain and Kumar, 2015; Ji, Chun, and Geller, 2013; Ji, Chun, and Geller, 2016; Hong and Sinnott, 2018).

Regarding COVID-19, many academic researchers in various fields including NLP have carried out studies targetting this subject. For example, COVID-19 and AI[5] is one of the conferences that has been convened virtually to show how Artificial Intelligence (AI) can contribute in helping Public Health Organisations during pandemics. Verspoor et al. (2020) organised a workshop on NLP for COVID-19 in two parts with 57 accepted papers that incorporated analyses in several languages. There are also multiple studies in various journals, conferences, and workshops with different goals such as monitoring the spread of the disease by finding infected individuals, defining the infected locations, or assisting governments and public health organisations in measuring people's concerns resulting from the disease. To clarify, there are about 2,610 papers on pandemics on ACL Anthology since the start of COVID-19 [6]. NLP is now being used by organisations to gain access to the landscape of research articles related to the COVID-19 pandemic. There is further discussion on this work presented later in Section 2.5.

2.2 Arabic Natural Language Processing (ANLP)

2.2.1 Arabic Language

The Arabic language is classified as the fifth most spoken language in the world[7]. It is spoken by around 400 million people around the world and has more than 26 dialects. There are a variety of Arabic dialects, some formal and some informal (Habash, 2010). Modern Standard Arabic (MSA) is the standard formal version in the Arab world, and almost everyone speaks and understands it. MSA is based on Classical Arabic, the

[5] https://hai.stanford.edu/events/covid-19-and-ai-virtual-conference

[6] https://aclanthology.org/

[7] https://en.wikipedia.org/wiki/Arabic

language of Islam's Holy Book, the Qur'an (Biadsy, Hirschberg, and Habash, 2009). It is the language of academia, the news and media, as well as law and legislation. Furthermore, the majority of research and methods in NLP is focused on MSA. In contrast, Dialectal Arabic (DA) is an informal language for the conduct of everyday life, television shows, and social media. Arabic dialects are less closely related to Classical Arabic than MSA. The dialects of Arabic differ from one another and from modern standard Arabic (Alshutayri, 2018).

For a variety of reasons, researchers are becoming increasingly interested in working on Arabic language processing. In 2017, the number of Internet users in the Arab world grew to 180 million[8]. The volume of Arabic textual data on the Internet has increased in the last decade, and the Arabic tools for browsing the web have improved. There is only one language, despite the fact that dialects vary from country to country and even within the same country.

2.2.2 Challenges in Arabic NLP

With such a large number of Arabic speakers, the focus on processing this language is rapidly growing. The Arabic language is written from right to left with 28 different characters. A big challenge for the Arabic alphabet is that letters shift their shape in accordance with their location. For instance, the letter kaff (ك), K in English, looks like (كـ) at the beginning, looks like (ـكـ) in the middle, and is formed like (ـك) at the end. The word originality was another complication, with 85% coming from tri-demanding roots. Therefore, Arabic is extremely derivational and inflectional, which results in morphological analysis being a complicated task. Arabic, like Chinese, Japanese, and Korean, has no capitalization. Since Arabic dialects are not standardised, their spelling is not always clear. It is important to emphasise an internal consensus as such must be used as normative orthodoxy in any approach to dialects. In other words, Arabic lacks a common orthography, resulting in a wide range of spelling variants. In addition, short letters are not represented orthographically in modern standard Arabic, which necessitates a high level of homograph resolution and word sense disambiguation. Arabic is a pro-drop language, as in Italian, Spanish, Chinese and Japanese, meaning that the topic pronouns can be deleted subject to recovery. In Arabic, synonyms are treated differently. A phrase may have up to seven synonyms, but the substitution of each of these terms is context-dependent, which may result in the extraction of redundant sentences. (El-Haj, 2012; Farghaly and Shaalan, 2009; Habash, 2010).

[8] `https://www.internetworldstats.com/`

Regarding our studies, the term (ڭورونَا), corona, can be found in many forms such as (كرونَا), (متكرون), and (كرونه). ALso, (انفلوَانزَا), Influenza, may appear in these expressions (فلو), (متفلوز), (حمَّى), and (محموم). All these words need to be included in our analysis research.

In summary, ANLP systems need to take into account the specific characteristics of the Arabic language in order to produce adequate results. There are some factors that have hindered the development of Arabic NLP in comparison to English and other European languages. These aspects include: the lack of capitalisation which led to difficultly of distinguishing proper names, words, and abbreviations; the need for complicated morphological rules in parsing text because vowels are absent in Arabic text which makes it difficult to know the meaning of the word; and morphological analysis is a difficult challenge since Arabic is heavily inflectional and derivational (El-Haj, 2012; Farghaly and Shaalan, 2009; Kanan et al., 2019).

2.2.3 ANLP Tools and Applications

ANLP has grown in popularity in recent years, and a number of cutting-edge technologies have been developed for a variety of applications, including machine translation, information retrieval and extraction, speech synthesis and comprehension, localization and multilingual information retrieval systems, sentiment analysis, text to speech, text classification and clustering and tutoring systems. These applications were designed to address many difficult questions pertaining to the Arabic language's meaning and structure (Farghaly and Shaalan, 2009; Habash, 2010; Kanan et al., 2019).

In terms of the complexity of Arabic morphology, ANLP researchers were successful in addressing this challenge, but further development is still needed. They face some problems when dealing with Arabic text. One difficulty is whether translations of Arabic words contain translated and transliterated words that do not adhere to a consistent set of rules on the way they are spelled in general. For example, a named entity like the city of Manchester could be spelled as (مَانشستر), (منشستر), (مَاشستر), (مَأَنشسطر), and(مَانشصتر). Another issue is the absence of a large corpus of Arabic-identified entities, which would have aided rule-based as well as statistical named entity recognition schemes. A third constraint is that, given the particular features of the Arabic language, NLP methods for western languages cannot be readily adapted to Arabic (Farghaly and Shaalan, 2009).

Researchers in ANLP use multiple methods to deal with Arabic morphological analysis to achieve the goal of analysing the text automatically without intervention from humans (Kanan et al., 2019). These techniques include:

Tokenization: is the process of dividing text into separate words called tokens by using delimiters like words punctuation and numbers (Alhanjouri, 2017).

Normalization: is the way to unify the form of the Arabic character due to the fact that one letter may have different shapes. For example, (أ ، آ ، إ) are converted to (ا) (Darwish, Magdy, and Mourad, 2012).

Part Of Speech Tagging (POS): is the method of tagging words with their grammatical type depending on the location and appearance in the sentence such as noun, verb, adjective, and adverb (Habash and Rambow, 2005).

Stemming: is determining the morphological stem of a word. In other words, it converts the word to the root by removing its affixes, which are prefixes, infixes and suffixes (Froud et al., 2010).

Stop Word Removal: is a method used to filter the text by removing the stop words in the language in order to improve results. In Arabic, they can be pronouns and conjunctions (Alhanjouri, 2017).

Named Entity Recognition (NER): is the process of extracting information and knowledge form text such as Names, Locations, Organisations, Phone numbers, Time, and Date (Ghosh, 2009).

Over the years, ANLP researchers have managed to enhance these tools because they are essential first steps in most natural language applications although the morphology of Arabic is seen as complex (Abuleil, Alsamara, and Evens, 2002; Al Ameed et al., 2005; Darwish, H. Hassan, and Emam, 2005; Khoja and Garside, 1999; Sawalha and Atwell, 2010). However, there is still a need to improve existing solutions to these challenges in order to adapt many applications of analysing Arabic corpora with accurate results. A study by Obeid et al. (2020) is a recent example of providing open source tools for ANLP in Python such as pre-processing, morphological modeling, dialect identification, named entity recognition and sentiment analysis. Although it achieved around 90% on the accuracy of some tasks, it needs to be further enhanced by incorporating additional datasets and models to accommodate additional dialects. In addition, the systems for identifying dialects were not trained using Twitter data, which present more challenges such as misspelling and the use of slang terms.

2.3 Analysing Social Media Data

2.3.1 Twitter Data Source

Twitter [9] is a microblog platform where posts contain less than 280 characters and additionally may contain some user information. There are 316 million monthly active users with 500 million tweets per day according to statistics from Twitter [10]. Twitter is one of the popular tools that can be used for an academic research. A study performed by (Ahmed, P. A. Bath, and Gianluca Demartini, 2017) explained some reasons for using Twitter in academic research: the accessibility feature compared with other social media platforms, the relatively simple method of finding conversations, the facility of collection due to strong hashtags, and the attention that is given to the platform by the mainstream media may attract more research.

On the other hand, there are many challenges when using Twitter as a data source. One of the challenges is finding the relevant tweets from an exceedingly large dataset (Doan et al., 2018). Another challenge is the false, noisy, or biased information that tweets may contain (Aramaki, Maskawa, and Morita, 2011; Hong and Sinnott, 2018). Moreover, tweets need more effort in preprocessing steps as a result of people using informal language when they tweet especially for the Arabic language as there are many different dialects. Finally, a set of Twitter terms and conditions that may obstruct the research while analysing tweets as explained in Section 3.2.

2.3.2 NLP for Social Media

In recent years, the use of social media has allowed users by their comments and posting on social platforms, to connect and express their thoughts, facts and awareness. There are several possible social networking platforms such as Facebook and Twitter, including the world's most used social channels. In social media, people choose to write messages that do not follow syntax, orthography or standard language, in their own languages.

NLP or Computational Linguistics is a field of artificial intelligence that is part of computer science. The objective of NLP is to make machines capable of understanding the language of humans. In other words, without the need for human interaction, NLP tools interpret written texts automatically. Although researchers face various challenges when analysing social media data, as described in the previous section (Section 2.3.1), there is a continually growing number of

[9] `https://www.twitter.com`

[10] `https://about.twitter.com/`

research ventures and publications leveraging social media. Different NLP core tasks are included in social media texts such as corpus annotating, part-of-speech tagging, linguistic pre-processing and normalization, information extraction, named-entity recognition, parsing, and multilingualism. These methods can be used for many applications like opinion mining, topic detection, sentiment analysis, geo-localization, machine translation, summarization, financial applications and healthcare applications (Farzindar and Inkpen, 2017).

2.3.3 Analysing Arabic Social Media Data

The researchers that analyse Arabic social media data encounter two different challenges: the Arabic language, Section 2.2.2, and social media data, Section 2.3.1 (Kanan et al., 2019). In addition to the application of NLP on social media, described in the previous section (Section 2.3.2), dialect classification is one of the important tasks in ANLP applied on social media data. The aim of this is to distinguish between different Arabic dialects because it is an important pre-processing step for other tasks, such as dialect-to-dialect lexicons, information retrieval according to the dialect, and machine translation (Alshutayri, 2018). There are also quite number of studies on opinion mining and sentiment analysis on Arabic social media data. The review by Kanan et al. (2019) summarised some papers on these fields. However, there is a lack of an established Arabic reference corpus and Arabic social media analysis is still in the early phases of production (Alshutayri, 2018; El-Haj, 2012; Farghaly and Shaalan, 2009).

2.4 Machine Learning

The method of programming computers to learn from training data (input) and display the results (output) is known as Machine Learning (ML) (Shalev-Shwartz and Ben-David, 2014b). It is a subfield of Artificial Intelligence that is focused on the premise that systems can learn from data, recognise patterns, and make decisions with minimal human involvement. ML is important because models can interpret larger, more complex data easily and automatically and provide quicker, precise results, even on the very large scale. It can be used in many application such as classification, prediction, extraction, and regression on either text, speech, or image.

2.4.1 Types of Machine Learning

ML is split into subfields based on learning tasks and results. The common types of algorithms are (Nassif et al., 2019):

Supervised Learning: where the machine is presented with a list of examples, along with the desired results set of values that the learner supplies, and the rule to be learned is generated from those examples. Examples of Supervised Learning: Regression, Classification, Decision Tree, and Random Forest.

Unsupervised Learning: this approach is unlike supervised learning, it trains the learning algorithm using an input dataset with no labelled outputs. Examples of Unsupervised Learning: Clustering, Anomaly Detection, and Dimensionality Reduction.

Semi-Supervised Learning: this approach combines supervised and non-supervised methods of learning, where there are several input dataset, some of them labelled (usually smaller) and the others not (usually larger). Examples of Semi-Supervised Learning: Self Training and Graph Based Algorithms.

Reinforcement Learning: where the past practise is used to teach the computer to make predictions. Examples of Reinforcement Learning: Monte Carlo and Heuristic methods.

Deep Learning: this approach can be defined as a sub-field of machine learning focused on multilevel algorithms, which provides a model that represents a complex relationship between data. Examples of Deep Learning: Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN).

2.4.2 Multi-Label Classification

Classification problem can be divided into three types: Binary, Multi-class, and Multi-label (Sorower, 2010; T. Li, Zhang, and Zhu, 2006). In machine learning, classification with a single label is a typical learning topic in which a variety of instances are learned from a group of disjointed classes, each of which is connected to an individual class label. The problem can be classified as binary classification, with two classes, or multi-class classification, when there are more than two classes, depending on the total number of disjoint classes. In contrast to these issues, multi-label classification requires instances to be assigned to several classes. In other words, it is used where one or more input values may be assigned to more than one classes. Multi-label classification is a supervised learning method and is more complex than binary and multi-class problems. Basically, there are three methods that can be used to solve the multi-label problem: Problem Transformation, Adapted Algorithm, and Ensemble approaches as described in the following (Madjarov et al., 2012):

Problem Transformation: is the process of simplify a multi-label problem to one or more single-label problems. One of the most well known examples is the separate classification of each label, as in the case of Binary Relevance and Classifier Chains and the separate class treatment of each label combination, as in the case of Label Powerset (Szymanski and Kajdanowicz, 2019).

Adapted Algorithm: is the approach that specifically manages multi-label data by extending basic learning algorithms (Tsoumakas and Katakis, 2007). For example, the k-nearest neighbors (kNN) classifier is expanded to the multilabel data by the MLKNN Algorithm.

Ensemble approaches: is the process of using multiple estimates or projections of function and combine them to create an efficient classifier. This assembly methods are one of the two groups of the multi-label methods, i.e. adaptation to algorithms and methods for problem transformation, which were previously described (Muller, 2018).

For each method, there are different techniques that can be used. The deep learning methods can also be used to solve this problem as explained in the next section, Section 2.4.3.

2.4.3 List of Machine Learning Algorithms

This section explains some of machine learning Algorithms that are used in this thesis for analysing text. These algorithms are chosen due to their powerful analysis in both Arabic text classification (Alshutayri, 2018; Ye Wang et al., 2017) and analysing social media for health surveillance (Aramaki, Maskawa, and Morita, 2011; Joshi, Karimi, Sparks, Paris, and C. Raina Macintyre, 2019b). The following are some descriptions of algorithms that have been used in our studies, applied in Chapter 4, Chapter 6, and Chapter 7, based on Supervised, Unsupervised, and Deep learning methods:

- Supervised Learning.
    - Binary and Multi-class Algorithms:
        - Logistic Regression (LR): is a statistical analysis that relies on probability and used for classification problems (Kleinbaum et al., 2002).
        - Random Forest (RF): It employs several different decision trees on multiple dataset subsamples. The average outcome is used as the model's forecast, which increases the model's overall statistical precision and combats overfitting (Shalev-Shwartz and Ben-David, 2014a).

        - ∗ Conditional Random Fields (CRF): It is a probabilistic system in which structured data are labelled and segmented such as strings, trees, and grids (Sutton and McCallum, 2012).
        - ∗ Support Vector Machine (SVM): It classifies the data points using a hyperplane in an N-dimensional space, where N here is the number of features (Ayodele, 2010).
            - · Support Vector Classification (SVC): it uses kernal functions, which can be linear, nonlinear, polynomial, radial basis function, and sigmoid, for classification process (Platt, 1999).
            - · Linear Support Vector Classification (LSVC): it uses a linear kernel for the basis function (Platt, 1999).
        - ∗ Bayes' theorem
            - · Naïve Bayes (NB): It operates with the Bayes theorem and the assumption of no attribute dependency, which implies that any feature in a class is unrelated to any other feature in the class. The data is used to compute a likelihood by assigning numbers to the likelihood of each instance in the package (Shalev-Shwartz and Ben-David, 2014a).
            - · Multinomial Naïve Bayes (MNB): It calculates the conditional probability of a given word based on its occurrence in a class, taking into account the number of times the word appeared in several classes (Schütze, C. D. Manning, and Raghavan, 2008).
    - – Multi-label Algorithms:
        - ∗ Binary Relevance: It treats each label as a separate single class classification problem (Sorower, 2010).
        - ∗ Classifier Chains: It treats each label as a part of a conditioned chain of single-class classification problems. It is useful to handle the class label relationships (Madjarov et al., 2012).
        - ∗ Label Powerset: It transforms the problem into a multi-class problem with one multi-class classifier that is trained on all unique label combinations found in the training data (Sorower, 2010).
        - ∗ Adapted Algorithm (MLKNN): It is a multi-label adapted K-Nearest Neighbors (KNN) classifier with Bayesian prior corrections (Madjarov et al., 2012).
        - ∗ Support Vector Machine with Naïve Bayes features (NBSVM): It combines generative and discriminant models together by adding NB log-count ratio features to SVM (S. I. Wang and C. D. Manning, 2012).

- Clustering.

  - K-means: It clusters data, which are unlabelled, by attempting to divide samples into n equal-variance sets, and the number of clusters must be defined (Shalev-Shwartz and Ben-David, 2014a).

- Deep Learning:
  - Bidirectional Encoder Representations from Transformers (BERT): It is a "pre-train deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers" (Devlin et al., 2018).
  - Transformer-based Model for Arabic Language Understanding (AraBERT): It is a pre-trained BERT model designed specifically for the Arabic language (Antoun, Baly, and Hajj, 2020).

We used LR, RF, MNB, and LSVS algorithms to classify the source of the tweets in Chapter 4. In Chapter 6, we applied LR, NB, and SVC algorithms to classify the tweets for rumour detection and K-means algorithm to explore the topics discussed on Twitter during the COVID-19. Since Chapter 7 include multi-label classification problem, we used Binary Relevance, Classifier Chains, Label Powerset, Adapted Algorithm (MLKNN), NBSVM, BERT, and AraBERT to identify infected people. We also used CRF algorithm to apply Named Entity Recognition in predicting the locations of affected people in same chapter, Chapter 7.

2.4.4 Feature Selection Methods

One of the most crucial stages in the classification process is feature collection. It is used to pick a subset of tokens or words that discriminate between classes and are present in the training set for use as features of text classification (Schütze, C. D. Manning, and Raghavan, 2008). In fact, choosing an appropriate feature would help to reduce the size of the effective vocabulary, increase the efficiency of training, and enhance classification accuracy. In this thesis, we used different feature selection methods with different goals and they are described in details in Chapter 4, Chapter 6, and Chapter 7.

2.5 Analysing Infectious Diseases Twitter Data

There has been a growing interest in Arabic language processing on social media data and Twitter in particular, but only a small part of this prior research is related to health and medicine topics as we show in this section. In contrast, a growing number of studies have analysed tweets for the public health information they contain in English and other languages such as Chinese and German (Charles-Smith et al., 2015). We summarise this research in what follows.

2.5.1 Analysing Health Related Arabic Twitter data

Before the COVID-19 pandemic, very few papers have described the use of Arabic tweets for studies related to health and medical topics. By 2020, Arabic researchers have begun to focus more on contributions in the Artificial Intelligence area especially in the NLP field. Here, we discuss some previous work on analysing Arabic social media data about general health topics and the work on the COVID-19 pandemic, since there are no studies related to other infectious diseases.

A survey performed by Alsobayel (2016) concluded that Twitter is the most frequently used by health care professionals in Saudi Arabia for the aim of professional development. Although the study was limited to health professionals in Saudi Arabia, it proved that social media platforms contain valuable information related to health. In (Alnemer et al., 2015) and (Albalawi, Nikolov, and Buckley, 2019), they analyse tweets in order to find the accuracy of the health information in Twitter and measure the rate of the tweets that contained false information, which is 51% in the former study and 31% in the latter one. The second study also tries to construe the features that affect the trustworthy of these tweets such as time to tweet, user linguistic and user popularity.

Several Arabic Twitter datasets related to COVID-19 have been recently published, for example (Alqurashi, Alhindi, and Alanazi, 2020; Haouari et al., 2020a; Qazi, Imran, and Ofli, 2020). Tweet ID's only can be found in these studies due to Twitter restrictions, as described in Section 3.2. They also explained the way of collecting tweets such as time period, keywords, and software library used in the search process and summary statistics for the collected tweets. Finally, they provide some suggestions for future work on these datasets, including: detecting misinformation, recognizing public opinions, monitoring a disease, understanding people's needs, and identifying knowledge gaps.

Hamoui, Alashaikh, and Alanazi (2020) analysed the tweets taking about COVID-19. They used the Non-negative Matrix Factorization (NMF) method to detect the topics and found the tweets' most popular unigrams, bigrams, and trigrams. While Alqurashi, Alashaikh, and Alanazi (2020) identified superspreader information during COVID-19 in order to understand the public reactions and impact of the information spreading. They applied PageRank and the Hyperlink-Induced Topic Search algorithms in their study and found Twitter influencers in the Arabic content of Twitter.

A study by Mubarak and S. Hassan (2020) provided a manually annotated dataset related to COVID-19 labelled with 13 classes. It outlined annotation instructions, analysed dataset, and developed classification models using different machine learning algorithms. In (Elhadad, K. F. Li, and Gebali, 2021), the authors presented a COVID-19 bilingual (Arabic/English) Twitter dataset that was automatically annotated to detect misleading Information. For building a credible ground truth database, they used the government websites and the shared official accounts as well as any relevant details on them. The dataset would then be labelled with different machine learning algorithms and various techniques of extraction.

2.5.2 Analysing Non-Arabic Twitter data

In contrast to Arabic, there is a considerably greater body of work in English that uses Twitter for studies on health and medicine-related topics. Previous studies have proved that Twitter contains valuable information on public health (Aramaki, Maskawa, and Morita, 2011; Breland et al., 2017; M. Paul and Dredze, 2011; Sinnenberg et al., 2017). These findings demonstrate the utility of NLP methods for extracting new knowledge from social media for public health analysis and supporting health informatics hypotheses. Here, we outline some studies relevant to past epidemics like Ebola, influenza, and measles as well as the COVID-19 pandemic.

Since 2011, researchers have been analysing tweets related to infectious diseases. M. Paul and Dredze (2011) and Aramaki, Maskawa, and Morita (2011) provided an analysis of tweets related to Influenza to detect infected people. The difference between the two studies is the first one concentrates on the distribution of words while the second one uses a sentence labelling system which classifies a tweet as negative when a person or surrounding person has flu in the present time.

The authors of (Arze et al., 2011; Ji, Chun, and Geller, 2012; Ji, Chun, and Geller, 2013) created the Epidemics Outbreak and Spread Detection System (EOSDS), which is an example of a platform that visualises Twitter users' worries over real health issues. On the first study they analysed tuberculosis, listeria, influenza, swine flu, and measles of tweets that contain disease name. Then they developed the system by making two-step sentiment analysis: first, they classify health tweets into personal tweets versus news tweets. Then they identify the tweets with negative sentiments that measure the degree of public's health concern over a disease. In the last study, they extended the system by analysing more diseases, which are 12 diseases including infectious diseases: listeria, influenza, swine flu, measles, meningitis, and tuberculosis; four mental health problems: Major depression, generalized anxiety disorder, obsessive-compulsive disorder, and bipolar disorder; one crisis: Air disaster; and one clinical science issue: Melanoma experimental drug.

In (Jain and Kumar, 2015), Influenza-A (H1N1) disease was analysed using Twitter data to retrieve the information that is important for health surveillance. Two approaches were used for selecting the keywords in the collection process, which are repeated weekly on specific period of time: (a) Keywords from tweets (b) Keywords from RSS feeds. The study presented a tool for using Twitter to monitor the H1N1 outbreak in India from February 1 to March 1, 2015.

The study by Takeuchi et al. (2017) looked at the frequency of disease-related terms like “nausea”, “chill”, and “diarrhoea”, and also used regression to predict the number of patients using Twitter data in Japan. It included only the tweets that have GPS information to compare the number of patients with the National Institute of Infectious Diseases (NIID).

Other researchers have combined Twitter data with Google Trends data for tracking the spread of infectious diseases like flu, chickenpox and measles (Hong and Sinnott, 2018). A study by Ahmed, Peter Bath, et al. (2018) analysed the Twitter data from infectious disease, swine flu and Ebola, outbreaks. The study developed new insights into how users respond during infectious disease outbreaks and reflects on users’ response in association with the sociological concept of the moral panic. They also suggest to examine the tweets in other languages. Table **??** summarizes the difference between some studies on social media for health surveillance in the past. Many of them used the name of diseases as keyword.

As far as COVID-19 research is concerned, according to (Shuja et al., 2020) open source COVID-19 datasets had been produced that contain medical images or textual such as reports, social media, and scholarly articles. (Banda et al., 2020) is an example of a large dataset in English, French, Spanish, and German languages. Researchers would be able to use this dataset to carry out a variety of studies on the emotional and behavioural reactions to social distancing interventions, as well as the detection of origins of disinformation and the stratified assessment of feeling against the pandemic in near real time as suggested in the paper.

Further research about Arabic and Non-Arabic COVID-19, that are more specific to our studies, will be discussed in Section 6.2 in Chapter 6 and Section 7.2 in Chapter 7.

Paper	Diseases	condition of tweets
Aramaki, Maskawa, and Morita (2011)	Influenza	a tweet person or surrounding person have flu and tense.
Ji, Chun, and Geller (2012)	tuberculosis, listeria,influenza, swine flu, and measles.	tweets contain disease name.
Ji, Chun, and Geller (2016)	12 diseases including infectious diseases, mental health problems, crisis, and clinical science issue.	two-step sentiment analysis: first, they classify health tweets into personal tweets versus news tweets. Then they identify the tweets with negative sentiments.
Jain and Kumar (2015)	H1N1	Tweets contain keywords from tweets and from RSS feeds and collecting tweets weekly on specific period of time.
Takeuchi et al. (2017)	Infectious Gastroenteritis	tweets with keywords and have GPS information.
Ahmed, Peter Bath, et al. (2018)	swine flu and Ebola	keyword with time interval selected from Google due.
Hong and Sinnott (2018)	flu, chickenpox and measles	collect all tweets then extracting illness-related tweets related to infectious diseases.

Table 2.2: Overview of past studies on social media for health surveillance

2.6 Identifying Geo-location from Tweets

The study of (Burton et al., 2012) identified the different forms of Twitter location information that could directly found when collecting tweets through Twitter API. This information could be: (1) the accurate tweet-related GPS coordinates, (2) the local GPS-coordinates (e.g. cities or metropolitan areas), (3) the free-text location information specified in the user account summary, and (4) the time zone. However, less than 1% of users provide these information in their profiles because they are an optional fields (Mahmud, Nichols, and Drews, 2014). Therefore, previous research uses tweet textual content to predict the location of the user (Khanwalkar et al., 2013).

Identifying locations in text is a part of Named Entity Recognition (NER) which is an NLP task of extracting information to recognise and classify information listed in predefined categories such as personal names (PER), organisations (ORG), locations (LOC) and dates and times (DATE). There are two dimensions of the previous research that need to explained here: Arabic Named Entity Recognition and Arabic Named Entity Recognition in Twitter.

2.6.1 Arabic Named Entity Recognition

Rule-based, machine learning-based, deep learning-based, and hybrid techniques are all used in NER methods. These methods are applicable to Arabic, though particular challenges of Arabic, as described in Section 2.2.2, such as lack of capitalization, nominal confusability, agglutination, and the absence of short vowels exist (Shaalan and Oudah, 2014; Zirikly and Diab, 2015). The survey by Shaalan (2014) summarised the advancement of Arabic NER research. It also illustrated the Arabic linguistic resources, methods used in Arabic NER, and evaluation metrics.

The study of (Salah and Zakaria, 2017) focused on machine learning research progress in ANER and compared linguistic resources, type of object, region, process, and results. It concluded that many reports had concentrated on supervised ANER machine learning studies, with little attention paid to semi-supervised studies, while the unsupervised approach is yet to be developed. It proved that most ANER machine learning structures concentrate on MSA with no regard to colloquial Arabic. Moreover, the machine learning NER research for MSA texts concentrated on few NEs and even fewer domains, whereas other topics, like like criminal history, religion, drugs, and sports were very rarely studied.

Recent advancements have shown that deep learning outperforms conventional NLP methods for different language processing activities, including NER. Theoretically, deep learning is expected to achieve better results by using an unending amount of training data, but in practise, data in the form of text is needed for good results in the Arabic language. Deep learning based on Arabian word embedding capturing syntactic and semantic connections between the terms will tackle the issue of Arabic NER. Recently, there have been some studies that applied deep learning on Arabic NER such as (Al-Smadi et al., 2020; Helwe and Elbassuoni, 2019).

2.6.2 Arabic Named Entity Recognition in Twitter

Much of the Arabic name entity recognition (NER) analysis is concerned with the task of Modern Standard Arabic (MSA), but it becomes much more important to study this challenge in the social media data. However, in social media data, there are more difficulties like Arabic dialects and informal words. In addition, many NER systems rely mostly on gazetteers that, due to inherent low coverage, can be more difficult in social media processing contexts (Zirikly and Diab, 2015).

The goal of the survey study in (Ali et al., 2020) was to produce a literature analysis of the numerous articles written in the field of NER on social media, focusing on the differences between the Arabic dialect and the English language. It demonstrated that the number of Arabic NER studies is still inadequate and very small. This is attributed to a variety of factors, including a scarcity of tools and resources, for example, an annotated dataset, gazetteers, and the Arabic language's complexity. Much of this necessitates several attempts to expand, diversify, and expand datasets in order to boost the output ratings of the various approaches.

2.7 Chapter Summary

In this chapter, infectious diseases, ANLP, analysing social media data, and machine leaning algorithms are briefly discussed. The literature review is focused on the previous research on analysing health-related Arabic and non-Arabic Twitter data and identifying geo-location from tweets.

None of the studies discussed in Section 2.5 have analysed the tweets to support better understanding of the spread of infectious diseases in the Arabic language for any Arab country. Therefore, there are no datasets, or ontologies related to health surveillance in Arabic. Moreover, non Arabic studies used keywords from formal or published styles of language which do not represent social media language. Using slang terms may help to find negation and irony in tweets which may indicate humorous and fake tweets. Common problems include concepts having different surface forms between formal and informal styles, or professional and non-professional descriptions of disease terminology and symptoms, and Arabic dialect detection in non-standard settings. The study of (Joshi, Karimi, Sparks, Paris, and C. Raina Macintyre, 2019b) suggested that there is a need to improve the quality of the epidemic intelligence by mitigation of false information. It also mentioned that is important to include informal text such as social media in the medical ontologies which contain scientific/technical terms, and these are precisely the areas that we tackle in this thesis.

The following chapter presents a description of the data collection process. It includes an explanation of the rules of Twitter, the Ethical approval, the process of collecting, and some statistics of the collected data.

Chapter 3

Data Collection

3.1 Introduction

Academic researchers are currently one of the largest Twitter API user communities. They have analysed tweets that represent public conversations since 2006, when the Twitter API was established, on many different topics. These topics may be related to climate change, politics, religions, sports, business, health, social, nutrition, and education. More details about analysing Twitter data using NLP techniques were explained in Section 2.3. Moreover, Twitter and other social media channels are analysed by researchers for their innovations and solutions in order to make the world a safer place to live in. In this chapter, we explain how the collection process is performed including adhering to the rules of Twitter, ethical approval, methodology, some statistics of the collected data, and discussion of some future directions that can be done within the dataset.

3.2 Rules of Twitter

According to Twitter's developer website[1], Twitter allows users to collect tweets to use them for academic research with the Twitter API. They can collect tweets depending on a set of keywords, location, or language with various tools and libraries using different programming languages such as JavaScript, Python, and Ruby.

Over the years, Twitter has been made some improvements and updates in the Twitter Developer Policy to facilitate academic researchers carrying out their experiments using Twitter data. Nevertheless, there are some steps that need to be followed to access the tweets. First, the user needs to register on the Twitter

[1] `https://developer.twitter.com/en`

application[2]. Then, applying for Twitter API access by describing the project that the user wishes to use the data for. Finally, the user can receive the API keys, which are a consumer key, a consumer secret, an access token, and an access token secret, that are needed to access the data. It is important to know that these keys are private and not shareable with others.

However, prior to 2021, it has not always been easy via the developer platform for researchers to access the information they need. For example, there was a restriction on the historical data, in other words, the ability to collect tweets from the past 7 days only was provided for free, but beyond that required buying the historical tweets from data providers. Therefore, researchers needed to develop more imaginative solutions to find the right information. For instance, to collect historical tweets you could collect tweets either weekly or using the streaming API.

In January 2021, Twitter changed its policy in order to improve support for the academic researchers (Tornes and Trujillo, 2021). Twitter's API now allows researchers to access all data for free without restrictions. To clarify, academic researchers can freely access the full historical record of tweets, access to the higher levels Twitter developer platform for free, filter the data to minimize cleaning requirements, and use new technical and methodological guides that maximize the success of the studies.

Although the rules of Twitter have changed, it is still important to carefully collect and build the dataset for every infectious disease. Not all devices can collect all historical tweets and save them. In other words, academic researchers need computers with continuous internet connections, and enough memory and storage to perform this task. Our dataset contains up to 20 infectious diseases and it is organised across the collected time period starting before January 2021. In addition, researchers need to compare their work with others and data sharing for replicability purposes can be achieved by sharing Tweet-IDs.

There are some benefits and drawbacks of using Twitter as a data source as described in Section 2.3.1. Regarding Twitter's rules, there is still a consequential limitation of the Twitter dataset such as tweets repetition via retweets but this can be resolved in the pre-processing step. The more effective limit is the deletion of tweets or data loss by Twitter. Twitter users may change their minds and delete their own content which requires you to delete the tweets from your own analysis and leads to other researchers not finding any deleted tweets in order to replicate the prior analysis of it.

[2] https://twitter.com/

Another problem that may affect the Twitter corpus is bot generated text. There is an increasing amount of spam in social media posts especially Twitter. Between 9% to 15% of active users are classified as autonomous agents (Varol et al., 2017) and recently NLP researchers have studied multiple techniques to detect bots (Inuwa-Dutse, Liptrott, and Korkontzelos, 2018). However, Twitter have made extensive efforts to detect these accounts and stop them by providing an enforcement team that can discover the users that generate them although it is not an easy process[3]. Twitter also encourages people to report spam accounts that satisfy the rules defined by Twitter[4]. The detection and characterisation of Arabic bots is as yet completely unexplored (Albadi, Kurdi, and Mishra, 2019; El-Mawass and Alaboodi, 2016).

In our research, the most clearly related area is the fake news in the health domain as discussed in Section 6.5.2 in Chapter 6. For our studies performed in Chapter 4, Chapter 6, and Chapter 7, while we are doing the analysis and the annotations, we have manually considered the tweets, and identified some of them which may be written by fake accounts. Then we review the spam policy provided by Twitter since this helps in checking whether the user could be classified as a bot, and if so we remove them from the dataset and this helps to ensure that any remaining content is related to real accounts.

To sum up, Twitter messages have recently emerged as a hot topic in NLP. The study of (Yang, Ning, and Y. Wu, 2020) proved that the combination of NLP and Twitter has been used in a wide range of businesses, employing the use of NLP. The industry's reliance on NLP-based Twitter predictions is continuously increasing as its accuracy improves. Nevertheless, there is a need for a lot more research to be done on the combined area of the NLP and Twitter.

3.3 Ethical Approval

Although Twitter has an informed consent from users to share information, there is a need to obtain research ethics approval from the University especially for the health related topics (Ahmed, P. A. Bath, and Gianluca Demartini, 2017). Ethical approval for this study was obtained from Lancaster University on 21st June 2019[5]. The full application for this approval can be found Appendix A.1.

[3] `https://blog.twitter.com/common-thread/en/topics/stories/2021/the-secret-world-of-good-bots`

[4] `https://help.twitter.com/en/rules-and-policies/platform-manipulation`

[5] `https://www.lancaster.ac.uk/sci-tech/research/ethics`

3.4 The Process of Collecting

There is a lack of an available and reliable Twitter corpus in Arabic in the health area which makes it necessary for us to create our own corpus. We obtained the data using the Twitter API since September 2019 collecting tweets from around 20 infectious diseases keywords. These disease search terms were determined depending on the information in the Ministry of Health in Saudi Arabia[6]. We collected the tweets weekly since the Twitter API did not allow us to retrieve enough historical tweets before the terms of conditions changed as described above. We used the names of the infectious diseases as keywords in the collection process. Table 3.1 shows the names of the infectious diseases that represent the keywords of the collection process. We have published these datasets for academic use.

During the collection process, a new disease may appear and/or new names of diseases are used. For instance, COVID-19 variants appeared in many different names which led the WHO to rename them with Greek letters [7]. The Greek letters refer to variants first detected in countries like the UK, South Africa, and India. So, the UK variant is Alpha, the South African version is Beta, and the Indian version is Delta, for example. As explained by the WHO, this was done in order to simplify talks, as well as to assist remove some stigma from the names whereby countries could be seen to be responsible for spreading a particular variant.

Around the middle of 2021, there has been a lot of news about the “black fungus”, to the extent that talk about it topped social networking sites in Saudi Arabia, Egypt, and other Arab countries. Rumours and exaggerations about the “black fungus” spread through the blogs. Tweeters talked about a “new epidemic”, “everyone who catches it dies”, or “his eye is pulled out”, if he is lucky and manages to live. Mucormycosis, or so-called black fungus, is a disease that reappeared in May 2021 and affected patients with Covid-19, or who were recovering from it[8]. As a result of this disease, there is a need to update the Ontology explained in Chapter 5 and collect tweets from June 2021 when cases appear in Arab countries (Egypt) and people started to tweet about it.

[6] https://www.moh.gov.sa

[7] https://www.bbc.co.uk/news/world-57308592

[8] https://www.bbc.co.uk/news/world-asia-india-57027829,https://www.bbc.com/arabic/trending-57220319

No.	Infectious disease name in Arabic	Infectious disease name in English
1	(التهَاب السحَايَا) الحمَى الشوكية	Meningitis
2	الحصبة	Measles
3	الَانفلونزَا	Influenza
4	(الَايدز) نقص المنَاعة البشري	Acquired Immunodeficiency Syndrome (AIDS)
5	البلهَارسيَا	Bilharzia (Schistosomiasis)
6	الحصبة الأَلمَانية	German Measles (Rubella)
7	الحمَى الصفرَاء	Yellow Fever
8	الحمَى المَالطية	Malta Fever
9	(السل) الدرن	Tuberculosis (TB)
10	الطَاعون	Plague
11	القمل	Lice
12	الـليشمَانيَا	Leishmaniasis
13	المَالَاريَا	Malaria
14	النكَاف	Mumps
15	انفلوَانزَا الطيور	Bird Flu
16	حمَى الضنك	Dengue Fever
17	دَاء الفيل	Elephantiasis
18	دَاء الكلب	Rabies
19	فيروس الورم الحليمي البشري	Human Papillomavirus Virus (HPV)
20	فيروس زيكَا	Zika Virus
21	فيروس كورونَا	Coronavirus

Table 3.1: Names of the infectious diseases

3.5 Statistics of the Collected Data

Here, we expound some statistics about the data collected from September 2019 until August 2021, which is the date of writing this chapter. Beyond this date, the Twitter collection process is ongoing and we will continue to add to the collection to be published for academic research. Figures 3.1a to 3.4d illustrate the number of tweets collected weekly of infectious diseases ordered as described in Table 3.1 during the collection process period, which is from September 2019 to August 2021. More details about the number of tweets can be found within the dataset in Appendix A.2 and as a downloadable file for researchers in github[9].

We can see from the figures, Figure 3.1a to Figure 3.4d, that the top five infectious diseases that have the highest number of tweets are COVID-19, Influenza, Plague, AIDS, and Malaria. People tweet about COVID-19 and Influenza as a result of the pandemic as discussed in Chapter 7. Around 542,900 tweets talk about Plague with 102,036 tweets from one week starting by 24 March 2020 as shown in Figure 3.2d. The reason for this high number is the spread of the disease during this period in China and led to many people sharing information about it [10]. According to the WHO[11], the human immunodeficiency virus (HIV) remains one of the major public health problems. People use the web search as shown in Figure 3.4e and social media such as Twitter as displayed in Figure 3.1d to find news and/or information related to it. No indigenous instances of Malaria have been reported in the Middle East Region in recent years, hence the majority of the countries in the region are Malaria-free. However, Malaria cases are still being imported from several of these nations (Al-Awadhi, Ahmad, and Iqbal, 2021). As a result, people may have some moral panic from Malaria and share some information in Twitter.

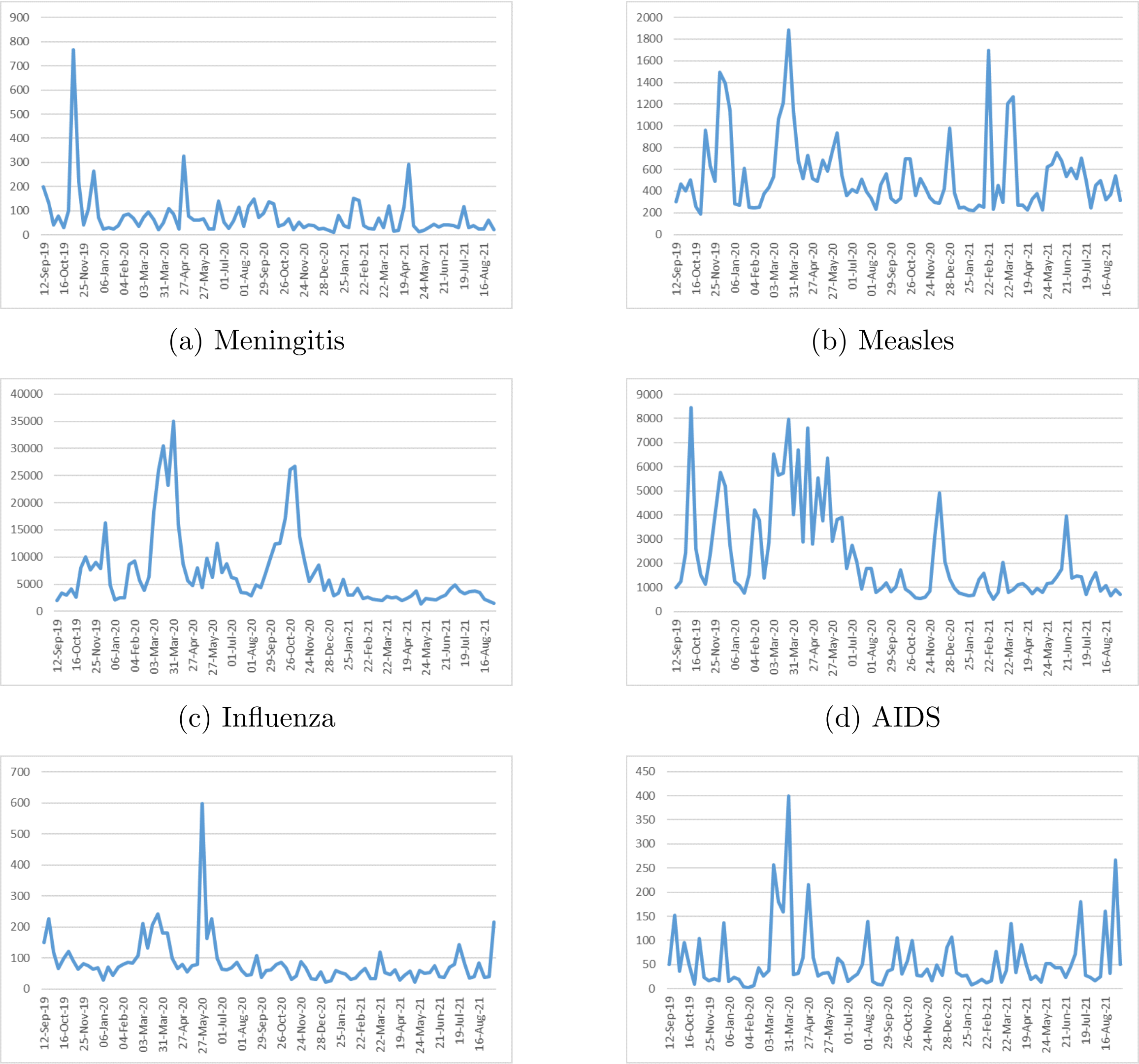
Before the COVID-19 pandemic, people tweeted about “Coronavirus” which was related to MERS-CoV , Middle East respiratory syndrome coronavirus[12] with only the disease name (كورونَا), Coronavirus. During the pandemic people use many different names including the informal one and the names of the Greek letters as described in the previous section, Section 3.4. In our dataset, people tweet using various names of COVID-19 as shown in Figure 3.4f. We can see that the term (كرونَا), which is an informal term of Coronavirus, was used in around 3,368,560 tweets and we proved the importance of these terms in Chapter 7. It is also essential to mention that people started using the names of the Greek letters after they were announced by WHO with 242,051 tweets containing the (دلتَا), Delta, term.

[9] https://github.com/alsudias/Arabic-Infectious-Diseases-Twitter-Dataset

[10] https://www.bbc.co.uk/news/world-asia-china-53325988

[11] https://www.who.int/news-room/fact-sheets/detail/hiv-aids

[12] https://bit.ly/3xCc7sr



(a) Meningitis

(b) Measles

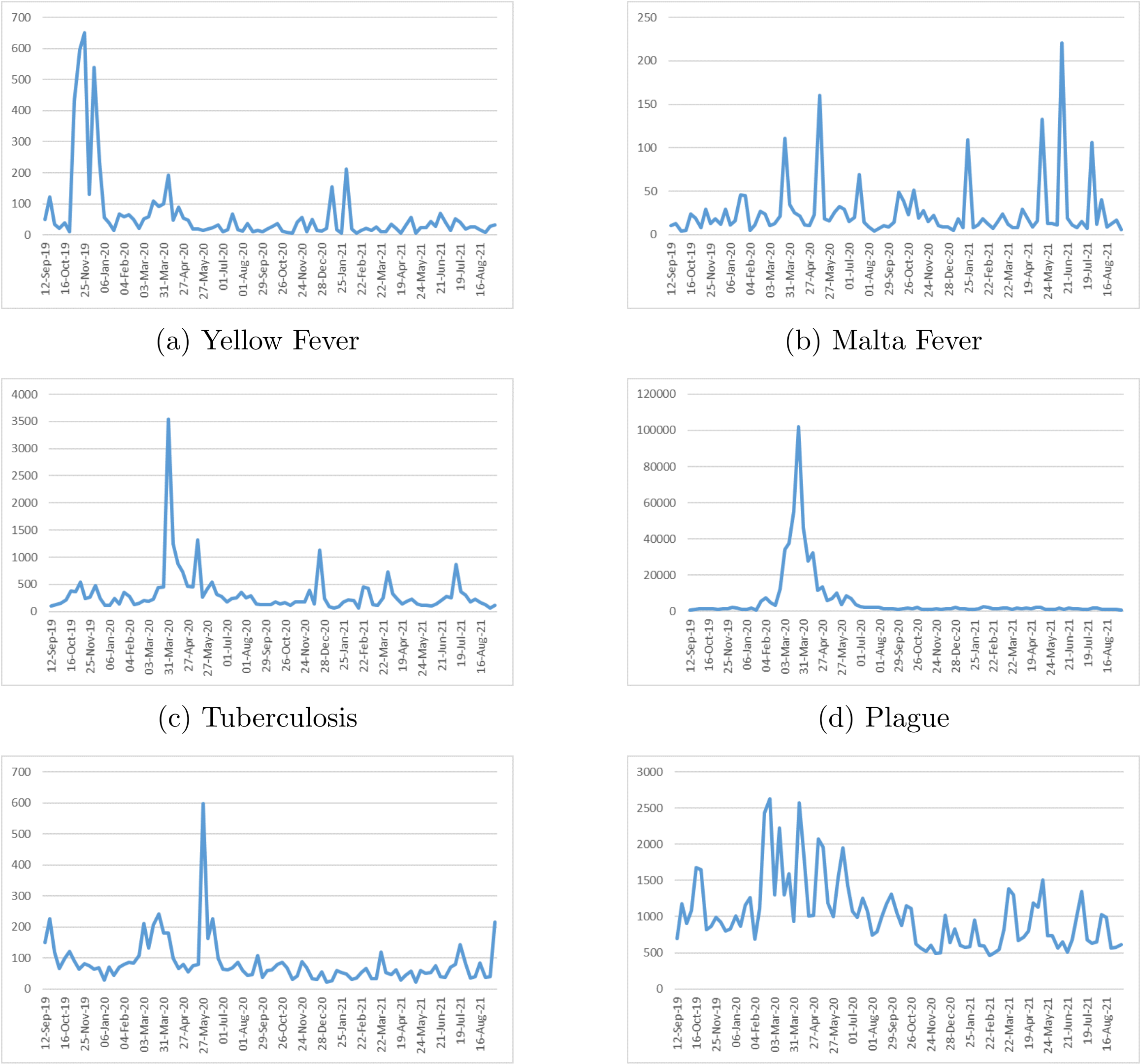
(c) Influenza

(d) AIDS

(e) Bilharzia

(f) German Measles

Figure 3.1: No. of tweets of different diseases



(a) Yellow Fever

(b) Malta Fever

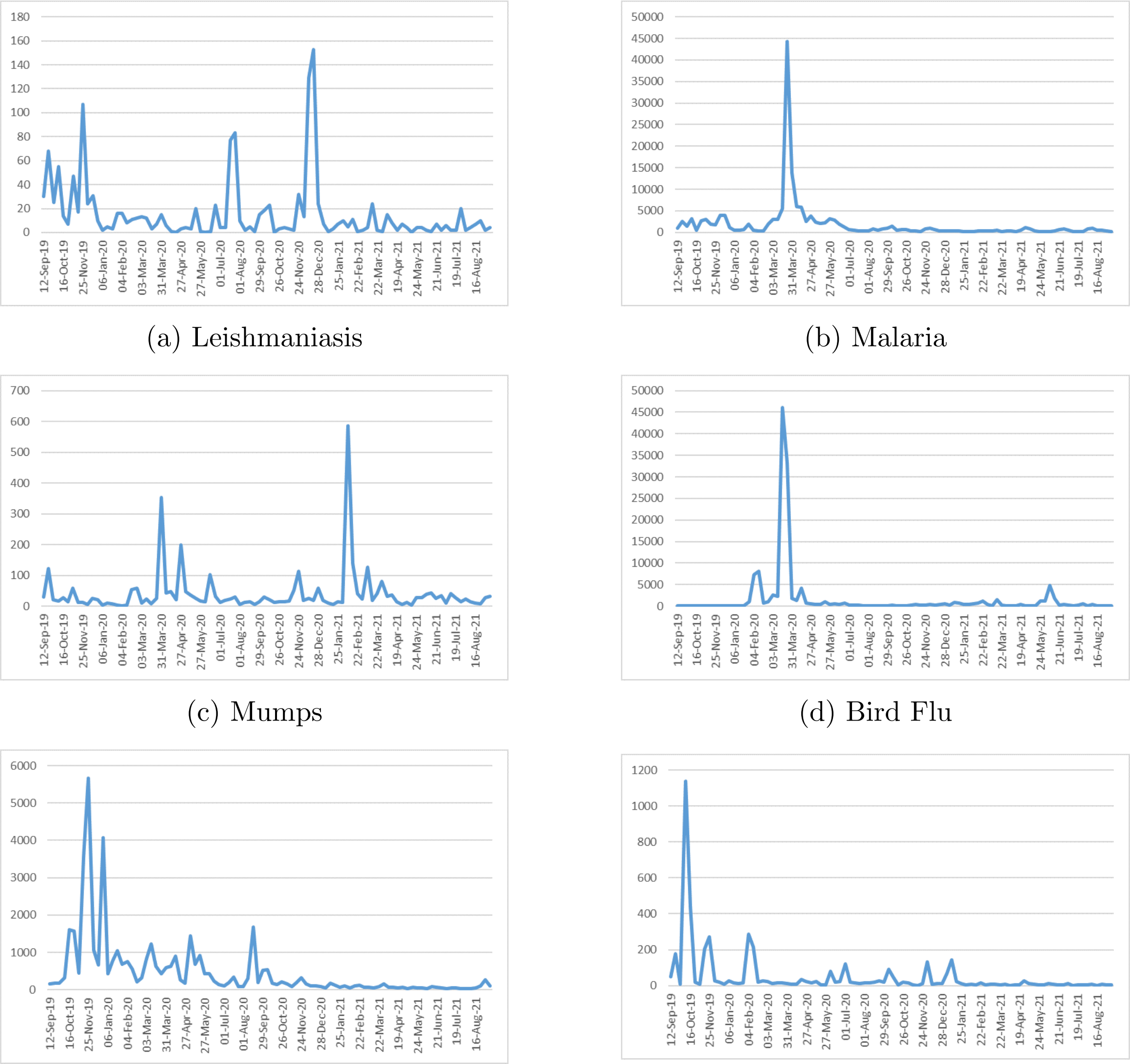
(c) Tuberculosis

(d) Plague

(e) Bilharzia

(f) Lice

Figure 3.2: No. of tweets of different diseases



(a) Leishmaniasis

(b) Malaria

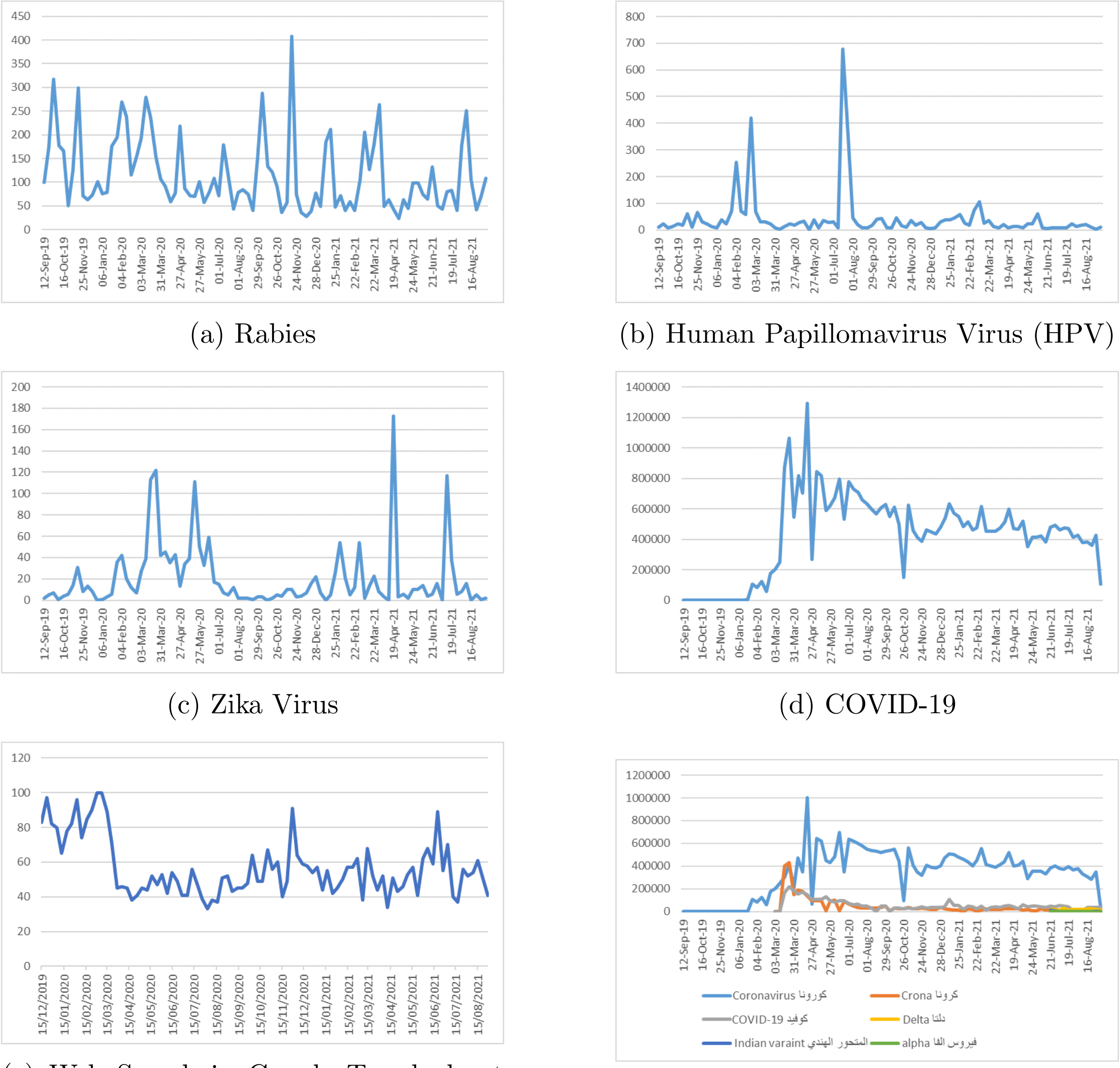
(c) Mumps

(d) Bird Flu

(e) Dengue Fever

(f) Elephantiasis

Figure 3.3: No. of tweets of different diseases



(a) Rabies

(b) Human Papillomavirus Virus (HPV)

(c) Zika Virus

(d) COVID-19

(e) Web Search in Google Trend about AIDS

(f) COVID-19 with different names

Figure 3.4: No. of tweets of different diseases

3.6 Future Directions of the Dataset

The Arabic infectious diseases dataset presented in this chapter is the largest Twitter dataset related to infectious diseases available to date. We believe it can foster social media health research in computer science in many application areas including the ones listed below.

- **finding the source of the tweet:** Knowing the person who has written the tweet is important to consider related to the truthfulness of and trust in the information. People are more likely to trust the tweet written by a reliable account but this is not always known by tweet content alone as studied in Chapter 4.

- **Analysing tweets to know the key topics:** During the pandemic or either epidemic, it is helpful to know what people are talking about to aid in gaining a better knowledge of a population's present and changing worries about the disease, as well as finding the best ways to safeguard people as discussed in Chapter 6.

- **Sentiment analyses of tweets:** As a result of using social media applications during disease spread, it is helpful to know people's opinions about many fields such as school, business, and government rules. For instance, analysing the impact of closing schools on student's education and health. Another example is finding some solutions resulting from closing shops like wearing a mask, social distancing, and buying online. Twitter data related to the Coronavirus epidemic illustrates how people, government institutions, and news organisations reported on the issue as shown in (Manguri, Ramadhan, and Amin, 2020). For an infodemiology study, Twitter data and machine learning methods may be used, allowing investigation into changing public debates and attitudes throughout the COVID-19 epidemic (Xue et al., 2020).

- **Detecting misinformation in tweets:** One of the most dangerous effects that health organisations faced during the pandemic is the false information that people passed between them via social media especially if the disease is new. We need to know this information to stop fake news by warning people about unreliable sources and false information. There is an example of this study in Chapter 6.

- **Analysing tweets for mental health:** One of the important fields that we need to focus on when diseases spread is mental health. People may feel panic from the news and information related to disease and/or bad implications from isolating or quarantine. For example, during the COVID-19 pandemic, the study by Koh and Liew (2020) looked at how people expressed loneliness on Twitter and the study by Su et al. (2020) compared the psychological effects of COVID-19 lockdown in China and Italy.

- **Monitor infected people:** When compared to traditional techniques, analysing tweets for infectious disease surveillance might be beneficial in reducing reporting lag time and providing an independent supplemental source of data. More details and proof can be found in Chapter 7.

- **Monitor self-isolating and quarantine:** There is a number of people that do not prefer to take the PCR (polymerase chain reaction) test when they feel some symptoms of COVID-19 to avoid self-isolating. Also, some people do not like to apply the quarantine rules after travel abroad. In this case, governments need to monitor those people who break the rules. As a result, people use social media to ask about treatment or express their emotions. With a small number of them satisfy the correct location in the profile, using either geo-location information or the content of social media can be useful to find these kinds of people although this could be considered an invasion of privacy.

- **Analysing tweets related to vaccination:** As a prominent source of health information, postings on Twitter may include information on vaccines, as well as sources, tone, and medical correctness. It is possible to gain insight about patient knowledge by analysing tweets, which may be used to develop teaching programmes (Love et al., 2013). These kinds of research may aid public health experts in better understanding the nature of social media's effect on vaccination attitudes and developing methods to combat the spread of misinformation (Luo, Zimet, and Shah, 2019).

- **Comparing different diseases:** Our Twitter dataset includes around 20 different infectious diseases. Researchers can compare different diseases at any period of time. For instance, a qualitative analysis of tweets on the Swine Flu and Ebola was provided in (Ahmed, Peter Bath, et al., 2018).

3.7 Conclusion

In this chapter, we showed the importance of building our social media corpus. We presented the process of collecting data as well as the explanation of the terms and conditions of Twitter. We provided some statistics of the data that has been collected. Finally, we demonstrated some future research that can be studied with this dataset, and several of these directions are followed in subsequent chapters. For each study in subsequent thesis chapters, Chapter 4, Chapter 6, and Chapter 7, that depend on this dataset, we made different pre-processing steps which are described in detail separately in these chapters. In addition, we used subsets of the dataset that represent some diseases within specific period of time.

Chapter 4

Information Sources of Arabic Infectious Diseases Tweets

4.1 Introduction

This chapter is about classifying information sources in Arabic Twitter to support online monitoring of infectious diseases. The objective of this chapter is to consider that it is important to assist reliability and trust judgements by taking into account the source of the information alongside the content of tweets. This chapter is derived from the published paper under the title "Classifying Information Sources in Arabic Twitter to Support Online Monitoring of Infectious Diseases" (Alsudias and Rayson, 2019).

People participate in the social web to express their opinions and provide information, which also gives researchers the opportunity to analyse those opinions across larger scale populations than is otherwise possible. One aim of this process is to summarise general opinions regarding international, national, or local events or themes in the huge amount of data available on the Internet. Twitter, one of the most popular tools for microblogging in real time, is used by people and organisations alike to share information on different topics, and these can include emergency and/or vital public health information. Due to its popularity as a communication platform, it can be difficult to distinguish reliable information from popular opinions or rumours and this is especially problematic in emergency or health-related scenarios.

Here, via the Twitter API we collected 1,266 tweets containing information about infectious diseases which we categorised into five types of sources: academic, media, government, health professional, and public. First, in order to validate the suitability of the groupings, two Arabic native speakers performed an independent manual labelling of each tweet into one of the types. The resulting inter-coder reliability was 0.84. Second, in order to see whether the grouping can be replicated on a much larger scale, we applied several machine learning models including Logistic Regression, Random Forest Classifier, Multinomial Naïve Bayes Classifier, and Linear Support Vector classifier. We evaluated the results using 10 fold cross validation for each model. The linguistic features we used to train the systems were selected via the best features based on univariate statistical test. The results show that the machine learning algorithms correctly classified tweets with up to 77.2% accuracy. We also prove that the bio of the tweet source is an important key that can be used in classification.

4.2 Related work

There has been a growing interest in Arabic language processing on social media data and Twitter in particular, but only a small part of this research is related to health and medicine topics as we show in this section. In contrast, a growing number of studies have analysed tweets for the public health information they contain in English and other languages such as Chinese and German (Charles-Smith et al., 2015). We summarise this research in what follows.

4.2.1 Arabic related research

Very few papers have described the use of Arabic tweets for studies around health and medicine topics. Alnemer et al. (2015) evaluated the correctness of health information on Twitter based on its medical accuracy. They found that 51.2% of tweets contained false information. The study showed the need for policies on using the social media for health care information. The study of (Alayba et al., 2017) used Twitter for sentiment analysis of health services. They collected unbalanced Twitter data and annotated it using three annotators by selecting the most frequent tags in each case. They used machine learning and deep neural networks techniques in their experiment and achieved 91% accuracy using support vector machines. A survey performed by Alsobayel (2016) concluded that Twitter is the most frequently used by health care professionals in Saudi Arabia for the aim of professional development.

4.2.2 English related research

In contrast to Arabic, there is a much larger body of work utilising Twitter in English for research on health and medicine related topics. The previous studies of M. Paul and Dredze (2011), Aramaki, Maskawa, and Morita (2011), Breland et al. (2017), and Sinnenberg et al. (2017) have proved that Twitter contains valuable information on public health. These studies show the power of NLP techniques for learning new information from Twitter for public health research and to support health informatics hypotheses.

The Epidemic Sentiment Monitoring System (ESMOS) is an example of tools that visualise Twitter users' concerns towards specific health conditions developed by Ji, Chun, and Geller (2013) in order to reduce the time of identifying peaks and ongoing monitoring of diseases required by public health officials. They also displayed a knowledge-based approach that utilises a medical ontology, an open source Disease Ontology developed by (Arze et al., 2011), to identify the occurrence of illnesses and to analyse the etymological expressions that provide subjective expressions and polarity of emotions, sentiments, conclusions, individual states of mind, etc. with an opinion classifier (Ji, Chun, and Geller, 2016).

In (Yepes, MacKinlay, and Han, 2015), pilot results were demonstrated in relation to future directions to investigate when using Twitter information for public health. Other research (Charles-Smith et al., 2015; M. Paul, Sarker, et al., 2016) has analysed social media articles in supporting public health in order to show the effectiveness of this strategy. These papers recommend a combination of social media with other techniques to measure disease surveillance and spread.

The goal of the study in (Sivasankari, Kavitha, and Saranya, 2017) was to produce a real time system for the prediction and detection of the spread of an epidemic by identifying disease tweets by graphical location. Good accuracy and quite diverse expressions were uncovered targetting health-related subjects (Doan et al., 2018). Although the results depend only on four months of Twitter data, the paper develops a useful approach to extracting cause-effect relationships from tweets.

Although the above studies use a variety of different approaches in NLP for showing how Twitter data can be used for public health applications including monitoring the spread of some diseases, they only consider the information content and do not attempt to study the trust or reliability of the sources of information. While social media users share information about disease outbreaks, symptoms, drug interactions, diet success, and other health behaviors (M. Paul, Sarker, et al., 2016; Yepes, MacKinlay, and Han, 2015), more than half of the tweets may contain false information (Alnemer et al., 2015). Hence, there is a clear requirement to filter this information in some way, and hopefully to study the ways in which we can reduce the noise of false information and to find tweets with more reliable information of a higher quality. A key first step to do this is to consider the source of the information since this helps the reader to determine how reliable the information is, for example by comparing a tweet on a specific topic by a health professional versus something similar via a member of the general public. Most previous research has focussed on the content of the tweets and not on the source of the information, hence we take a new approach in this chapter to combine the two elements together.

4.3 Data Collection

First, we collected 10,000 Arabic tweets via the Twitter API between December 2018 and February 2019. We defined the keywords related to infectious diseases. These keywords were generated by translating an English Disease Ontology, a medical Ontology, developed by Schriml et al. (2011) since we were unable to locate a suitable Arabic equivalent. Using the Disease Ontology allows us to find all terms related to infectious diseases with a set of synonyms (Ji, Chun, and Geller, 2016). The Twitter API does not allow us to retrieve enough historical tweets unless we know the user ID in advance. Therefore, we followed two strategies in collecting tweets:

- collect tweets containing words from the keyword list.
- from a suitable user account (discovered using the first step), collect all historical tweets, and filter them depending on the keyword list.

Next, we filtered the tweets manually by removing advertisements, spam, and retweets. After that, we devised Python scripts to clean the tweets by removing URLs, mentions, hashtags, numbers, emojis, repeating characters and non-Arabic words. We also automatically normalised the Arabic tweets and tokenised them. The author of the thesis first classified each of the tweets manually into one of the five groups described in Section 4.4. Second, we asked another Arabic native speaker to independently classify the tweets into the same five groups in order to calculate inter-coder reliability as described in Section 4.5.

4.4 Tweet Categorisation

Based on our initial manual reading of the tweets, we decided on five types of users to classify the tweets into, taking into account the various levels of trust that might be associated with each type. For each category listed below there is a small description with examples illustrated in Table 4.1.

Academic: the tweet is written by academic researchers in higher education. To illustrate, this could be a researcher who carries out studies about infectious diseases.

Media: the tweet is written for newspapers or magazines whether they are general media or health specific ones. In most cases it contains news about infectious diseases.

Government: the tweet is written by a user account that represents the government in some official capacity such as the ministry of health. It may include news, admonition, warning, or general information related to infectious diseases.

Health professionals: the tweet is written by doctors, nurses, or other health service practitioners. In other words, any person who is employed or trained in the health domain and writes the tweet on any information related to infectious diseases.

Public: the tweet is written by members of the general public. It may include information on infectious diseases, feeling sick, or giving advice to someone. Also, it may be written in many dialects since the people come from many Arab countries.

Table 4.2 shows the number of tweets in each category after filtering and preprocessing. The total number of tweets is 1,266 with only 56 tweets in the media category and 436 tweets in the public one. The reason for this is that there are few media accounts that tweet about infectious diseases whilst many members of the public tweet on the topic. In the other categories (academic, government, and health professionals), there are a relatively well balanced number of tweets which are 239, 258, and 277, respectively.

4.5 Manual Coding and Inter-coder Reliability

4.5.1 Annotation Process

The process starts with labeling the tweets by two Arabic native speakers, including the author of the thesis, who were provided with the guidelines detailed above. The classification depends first on the text in the tweet itself. If the tweet has an ambiguous classification, the annotator may need to look at the bio, a description written by the Twitter user, of the user tweet.

Category	Tweet in Arabic	Translated tweet to English
Academic	ورقَة علمية حديثة ترَاجع وتّنَاول فيروس الهربس وتَاثيرَاته الَاكلينيكية خَاصة علَى الَاطفَال ومرضَى ضعف المنَاعة	A recent scientific paper reviews the Herpes virus and its clinical effects on children and patients with immunosuppression.
Media	حَالة وفَاة بسبب عدوَى الكورُنَا في خميس مشيط	A case of death due to infection of the corona in Khamis Mushayt.
Government	يتوفر لقَاح الَانفلوَنزَا في مرَاكز الرعَاية الصحية الَاولية	The flu vaccine is available in primary health care centers.
Health Professional	لَا تستعمل قطرَت الَاذن من توصية مريض سَابق ، فعلَاج التهَابَات الَاذن الخَارجية البكتيري يختلف عن علَاج الَالتهَاب الفطري	Do not use ear drops from a previous patient recommendation. Treatment of bacterial external ear infections is different from treatment for fungal infections.
Public	العشبة العجيبة لعلَاج اكثر من مرض في عشبة وَاحدة وهي الزنجبيل	The wonderful herb to treat more than one disease in one herb is ginger.

Table 4.1: Examples of tweets in each category

4.5.2 Inter-coder Reliability

We used the Kappa Statistic to test the robustness of the classification scheme (Artstein and Poesio, 2008). The result showed that Cohen's Kappa score was **0.84** which indicates strong agreement between the two manual coders. Figure 4.1 shows the confusion matrix between the two annotators. The most disparate results between the two coders is between the academic and health professionals categories and this accounts for 10.9% out of the 16% total. This is caused by the similarity between the language of the two groups and the lack of explicit information in the bio to determine which category the tweet fits into.

Category	No. of tweets	No. of words
academic	239	3,696
media	56	601
government	258	3,602
health professional	277	6,123
public	436	4,963
total	1,266	18,985

Table 4.2: Number of tweets and words in each category

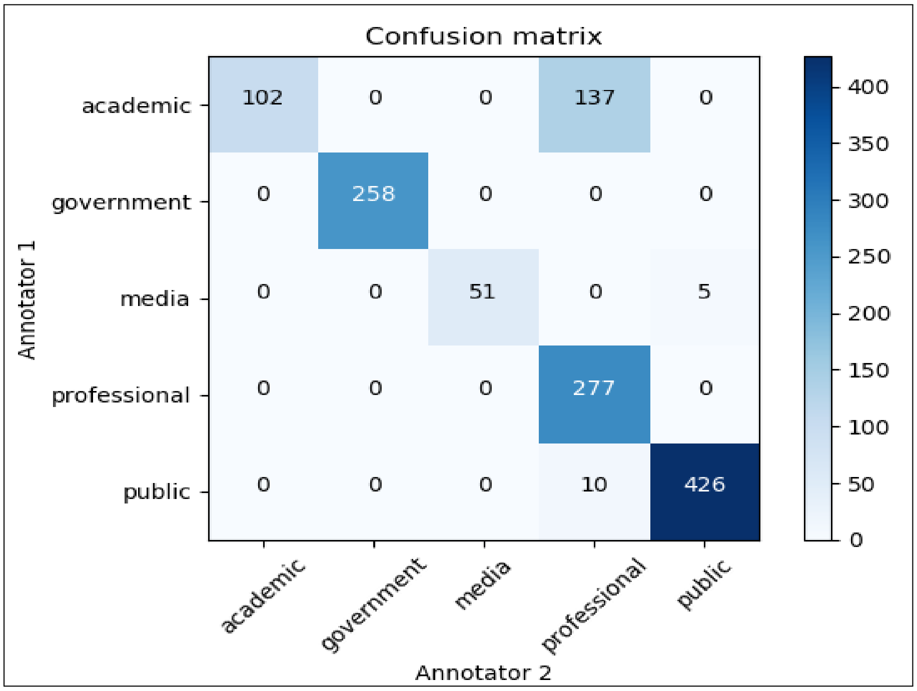


Figure 4.1: Heat map of confusion matrix between the two annotators.

4.6 Machine Learning Models

In our study, we used Python Scikit-learn 0.20.2 (Pedregosa et al., 2011) software and applied four machine learning models: Logistic Regression, Random Forest, Multinomial Naïve Bayes and LSVC: Linear Support Vector Classification. These algorithms tend to be the most used techniques in classifying social media health surveillance literature (Aramaki, Maskawa, and Morita, 2011; Joshi, Karimi, Sparks, Paris, and C. Raina Macintyre, 2019b) and Arabic text classification (Alshutayri, 2018; Ye Wang et al., 2017).

We used 10-fold cross validation to determine accuracy, splitting the entire sample into 90% training and 10% testing for each fold.

4.6.1 Machine Learning Features

A word frequency approach was used to extract the features from the processed training data after converting them to a matrix of token counts. We designed several features to be used in all four algorithms in order to obtain the best accuracy. These types of features were used in various previous work to distinguish the style of different individuals such as identifying user identity in Twitter (Keretna, Hossny, and Creighton, 2013; Green and Sheppard, 2013).

4.6.1.1 Feature selection

We used various techniques to select the best features automatically (Pedregosa et al., 2011):

- all features: counting unigrams, bigrams and trigrams and ignoring terms that have a document frequency strictly lower than two.
- selecting the best features: based on a univariate statistical test, keep 60% of the features that have the highest scores.
- selecting from a model: Random Forest Classifier is used to remove unimportant features.
- Using Stanford POST (Part Of Speech Tags) (C. Manning et al., 2014): Arabic POST is used to annotate the tweet with part-of-speech tags.

4.6.1.2 Balancing the data

Since the number of tweets is unbalanced across the types, we used two different techniques to re-sample them: under-sampling and over-sampling. We used RUS (Random Undersampling) for under-sampling and ROS (Random Oversampling) for over-sampling. RUS works by removing samples randomly from the majority classes while ROS generates more examples for the minority classes, which has less training data (Burnaev, Erofeev, and Papanov, 2015).

4.6.1.3 Using user bio as a feature

To further resolve the close ambiguity of some types such as academic and health professional, or government and media, we used the bio of the tweet user to provide further features in combination with the tweet text itself. Then we repeated the experiment, including preprocessing, feature extraction and selection, and applied machine learning algorithms, with the new text. Table 4.3 shows an example of tweets that may be classified into different classes. The first example, which is written by a health professional, can be classified as government because it seems to be providing advice from the government or as an academic as a result from their research.

Tweet in Arabic	Translated tweet to English	Category
التوصية بَاخذ تطعيم فيروس الحصبَه للّمسَافرين.	Recommend vaccination of the measles virus to travelers.	Health professional, government, or academic.
الربيعة: افتتحنَا عدة مرَاكز اضَافية لمرضَى الَاضطرَابَات و غيرهَا وتوسعنَا في برَامج صحة المرَاه خصوصَا و اطلقنَا مُوءشرَات لقيَاس ادَاء المرَاكز الصحية و رضَا المرَاجع	Alrabiaa: We have opened several additional centers for patients and disorders and others and expanded in women's health programs, especially we have launched indicators to measure the performance of health centers and the satisfaction of references.	Media or government.
بحث لثلَاثة أطبَاء سعوديـين يكشف علَاقة انزيم الكبد الدودية بَالتهَاب الزَاءيـده	A study by three Saudi doctors reveals the relationship of the liver enzyme to Appendicitis.	Media or academic

Table 4.3: Examples of tweets with ambiguous categorisation

4.7 Results and Discussion

Choosing the most frequent class (public) represents 34.4% of the data set, so this represents a simple baseline for our results. Table 4.4 shows the accuracy, F1-score, recall, and precision of the machine learning models on our training dataset. The Logic Regression classifier and Multinomial Naïve Bayes achieved the highest accuracy (76.0%), with higher recall (0.76) compared to the other algorithms. To evaluate the automatic classification, we compared the algorithms with different sets of features. Figure 4.2 illustrates the results we achieved running the four machine learning algorithms with different set of features. The highest accuracy (77.2%) is achieved by the Logic Regression classifier with a selection of the best features based on a univariate statistical test features. It also provided the best score on Random

Machine Learning Model	Accuracy%	F1-Score	Recall	Precision
LR	76.0	0.74	0.76	0.74
RF	65.1	0.66	0.68	0.71
MNB	76.0	0.75	0.76	0.75
LSVC	74.1	0.73	0.74	0.75

Table 4.4: Training results using all features
LR: Logistic Regression, RF: Random Forest, MNB: Multinomial Naïve Bayes, LSVC: Linear Support Vector Classification

Forest and Linear Support Vector classifications (68.2% and 76.1%, respectively) while selecting all the features reached 76.2% in multinomial NB classifiers. Selecting a model to remove unimportant features did not provide any better results via the four algorithms and POST features had the worst results.

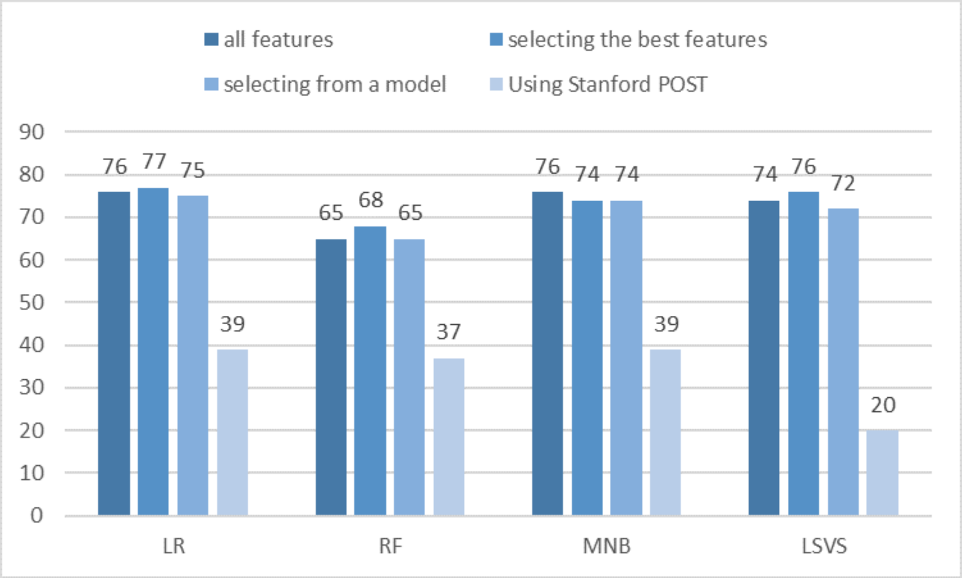


Figure 4.2: Effect of different features on classification.
LR: Logistic Regression, RF: Random Forest, MNB: Multinomial Naïve Bayes, LSVC: Linear Support Vector Classification

We also compared the normal data, which is the original unbalanced data, with the over and under sampled data to show the effect of re-sampling on the classification accuracy, and the results are shown in Table 4.5 and Table 4.6. We can see that there is a small enhancement of accuracy using an over-sampling method especially when selecting the best features and selecting features from a model in all four classifiers.

Machine Learning Model	all features	selecting the best features	selecting from a model
LR	76%	79%	77%
RF	62%	72%	72%
MNB	72%	78%	75%
LSVC	73%	78%	76%

Table 4.5: Effect of over-sampling data on classification

Machine Learning Model	all features	selecting the best features	selecting from a model
LR	73%	70%	66%
RF	53%	56%	49%
MNB	68%	67%	61%
LSVC	63%	61%	67%

Table 4.6: Effect of under-sampling data on classification

On the other hand, under-sampling the data achieves lower accuracy than normal data due to losing some data in the re-sampling process. Figure 4.3, Figure 4.4, and Figure 4.5 show the comparison between normal data, over-sampling, and under-sampling in all features, selecting the best features and selecting features from a model respectively. We can see that when applying all features, the normal data has the best result in all models expect Logic Regression which has the best result when using over-sampling data with accuracy 77.1% (Figure 4.3). However, over-sampling data with selecting the best features raises the accuracy between 2% to 4% above normal data in all models (Figure 4.4). It reaches 79.1% in the Logic Regression classifier and 78.2% in the Multinomial Naïve Bayes and Linear Support Vector Classification models. Moreover, the accuracy is highest when over-sampling data with selecting features from a model in all classifiers for instance, Random Forest classifier which increase from 65.1% to 72.2% (Figure 4.5).

Table 4.7 represents the accuracy of each model after combining the bio of the tweet user with the tweet text. In many of the results, the accuracy improves to **99.9%** which shows that the user bio has a very important role to play for classification. For instance, example number 2 in Table 4.3 can be classified as government after reading the bio of the Twitter user which is:
(تـتمثل رؤُية وزَارة الصحة في تحقيق الصحة بمفهومهَا الشَامل علَى المستويَات لـلفرد
و الأَسرة وَالمجتمع مع العمل علَى مسَاعدة المسنين و الأَشخَاص ذوي الاَعَاقة بمَا يمكنهم)

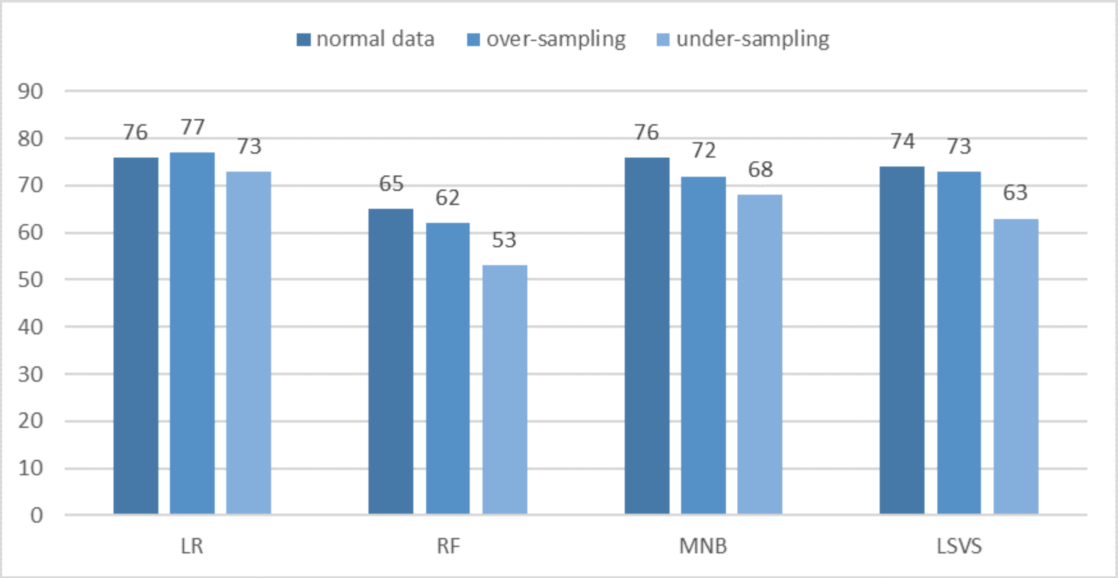


Figure 4.3: Effect of re-sampling data with all features on classification.

Machine Learning Model	Tweet text only	Tweet text with bio
LR	77.2%	99.9%
RFt	68.2%	99.9%
MNB	76.2%	99.6%
LSVC	76.1%	99.9%

Table 4.7: Effect of using bio of tweet user as feature on classification

which means in English: ”The vision of the Ministry of Health is to achieve comprehensive health at the individual, family and community levels while working to help the elderly and those with special needs”. There are some words in the example bio such as ministry which can help in the classification process. The very high accuracy of the classification can be explained as a result of the limited number of Twitter accounts in the study.

We performed an analysis using the Logic Regression classifier with a set of the best features based on univariate statistical test features in the heat map in Figure 4.6. The matrix shows 12 tweets from the academic category are confused with the government category. In Figure 4.7 representing the worst accuracy (65.1%) resulting from Random Forest with all features, we assess the degree of confusion between categories. The highest confusion across all categories is between the academic and government types.

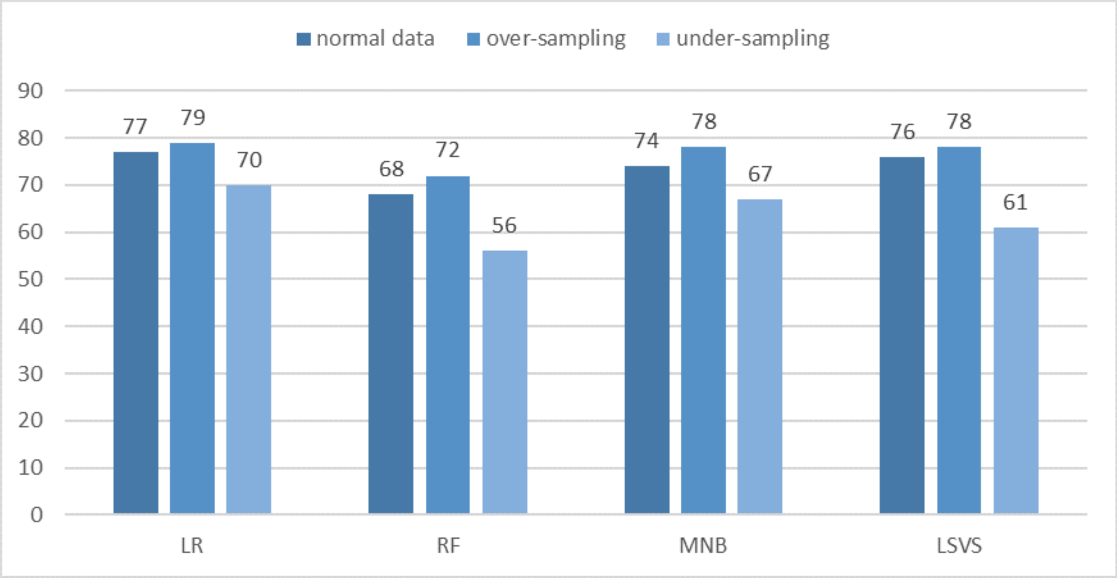


Figure 4.4: Effect of re-sampling data with selecting the best features on classification.

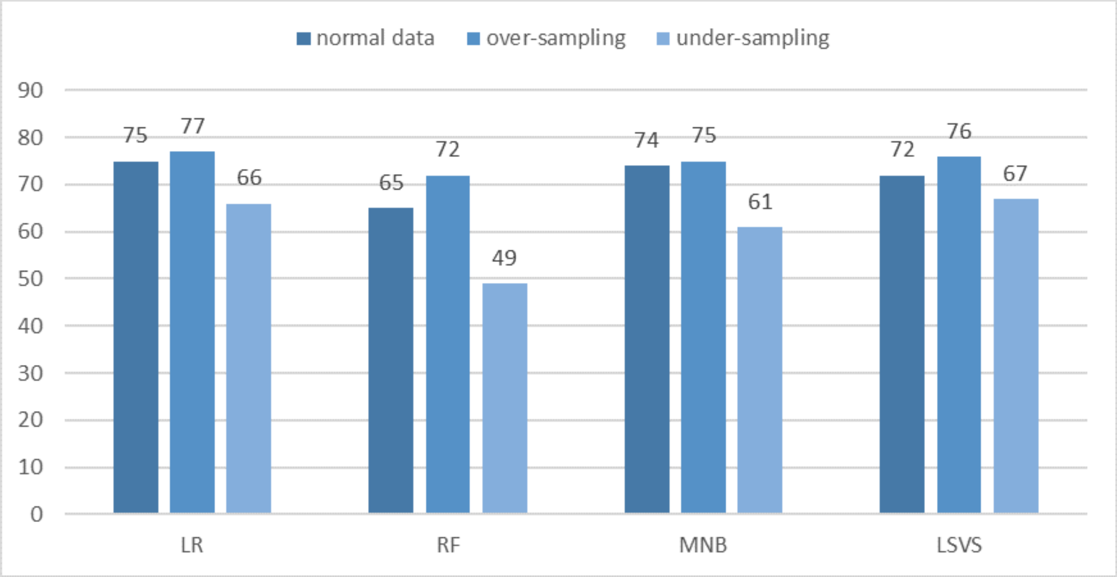


Figure 4.5: Effect of re-sampling data with selecting from a model on classification.

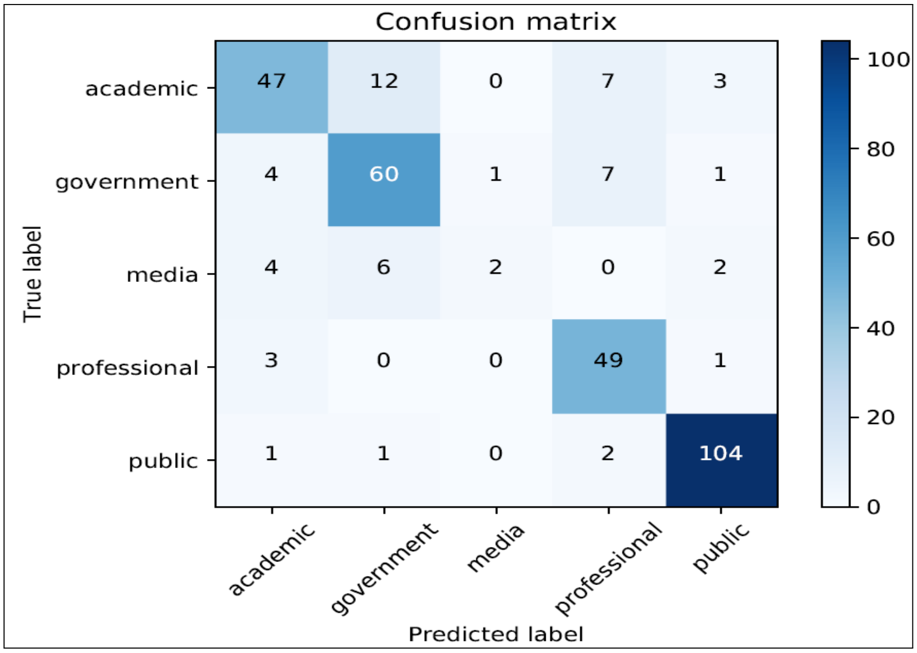


Figure 4.6: Heat map of confusion matrix of the Logic Regression classifier with selecting the best features based on univariate statistical test features.

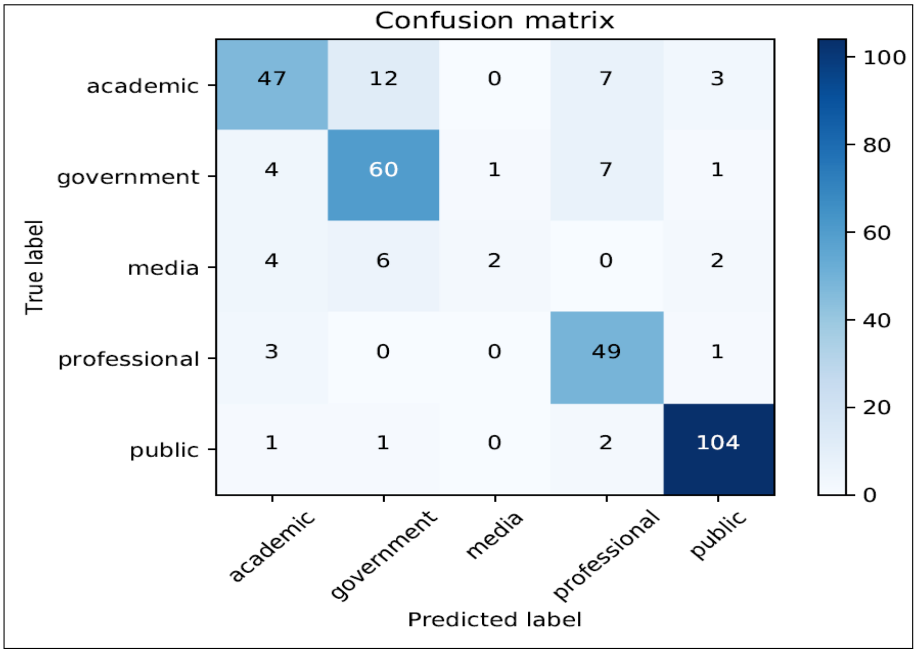


Figure 4.7: Heat map of confusion matrix of the Random Forest classifier with all features.

4.8 Conclusion

In general, social media can be useful as a source of health information. Many government organisations, academic experts, professions, and media outlets in the field of health and medicine release useful health information needed by the public. However, in emergency situations and fast moving scenarios, it is important to understand the veracity of information released in social media, in order to avoid acting on false information. Therefore, we study not only the content of tweets but also the source of the information in order to work towards determining the quality and reliability of public information. We use the bio of the Twitter user which contains key information in order to discover reliable accounts that can be trusted.

Here in this chapter, we introduced a new Arabic social media dataset for analysing tweets related to infectious diseases. The dataset of Arabic tweets has been manually classified into five categories: academic, media, government, health professional, and public, with good inter-rater reliability. We then used machine learning algorithms to replicate the manual classification. The results showed high accuracy on the classification task, and showed that we are able to classify tweets into the five categories with favourable accuracy on the tweet content itself, and highly accurately using the bio information from the user. The dataset, including tweet ID, manually assigned categories, and other resources used in this chapter are released freely for academic research purposes[1].

We save the model, which is useful for classifying tweets into five categories: academic, media, government, health professional, and public. We use Logic Regression model because it previously achieved the best accuracy (77%). We do this in order to predict the source type for each of the tweets later such as the prediction of the source of the COVID-19 tweets in Section 6.5.3 in Chapter 6.

To sum up, we found that different types of groups can tweet about infectious diseases and the public group use an slang words when they tweet. Therefore, there is a clear need to include the informal terms when studying the surveillance of the infectious diseases. Moreover, there is demand to involve not only the infectious diseases name as a keywords when collecting tweets but also all terms related to them such as symptoms, treatments, and infection ways. As a result, we need to build an Arabic Infectious Disease Ontology, which described in details in the next Chapter, Chapter 5, to help in monitoring of infectious disease spread via social media.

[1] https://doi.org/10.17635/lancaster/researchdata/303

Chapter 5

An Arabic Infectious Disease Ontology

5.1 introduction

This chapter is based on an extended version of the "Developing an Arabic Infectious Disease Ontology to Include Non-Standard Terminology" paper (Alsudias and Rayson, 2020c). It presents a new Arabic ontology in the infectious disease domain to support various important applications including the monitoring of infectious diseases spread via social media. This ontology meaningfully integrates the scientific vocabularies of infectious diseases with their informal equivalents. The overall goal of this ontology is to be a key source of information related to infectious diseases in Arabic. It can be used for many important applications in academia and the real world including Question Answering on the Semantic Web (AlAgha and Abu-Taha, 2015) and, in particular in our case, for online monitoring of health events (Collier et al., 2010).

Ontologies are formal representations of concepts and their relationships in machine readable form. Ontologies can be built manually, semi-automatically or automatically. Generally, they fall into two categories: Upper level Ontology and Domain Ontology. There are many existing ontologies in English and other languages but for Arabic there has been little recent research. Developing ontologies from Arabic text is a difficult process due to the lack of existing resources and the nature of the Arabic language which has complex grammatical, morphological and semantic aspects. Existing NLP algorithms for other languages cannot simply be applied directly on Arabic. In addition, the Arabic language has many dialects which may complicate ambiguities for computational understanding of the meaning of the words and reduce the accuracy of tools created only for modern standard Arabic.

Here, we used ontology learning strategies with manual checking to build the ontology. We applied three statistical methods for term extraction from selected Arabic infectious diseases articles: TF-IDF, C-value, and YAKE. We also conducted a study, by consulting around 100 individuals, to discover the informal terms related to infectious diseases in Arabic. We created the relations for infectious disease concepts manually. We reported two complementary experiments to evaluate the ontology. First, a quantitative evaluation of the term extraction results and an additional qualitative evaluation by a domain expert.

The chapter is organised as follows: Section 5.2 covers related background work on ontologies. Section 5.3 explains the requirement for an Infectious Disease Ontology in Arabic. The Ontology Engineering process includes Ontology Architecture, Data Gathering, Pre-processing, Term Extraction, Finding Synonyms, Adding informal Terms, Obtaining Concepts, and Defining and Updating Relations discuss in Section 5.4. Section 5.5 clarifies the Ontology Implementation. Two types of ontology evaluation are performed in Section 5.6. Section 5.7 states the conclusion.

5.2 Related Work

Building ontologies is a crucial part of the semantic web endeavour. In recent years, research interest has grown rapidly in supporting languages such as Arabic in NLP in general but there has been very little research on medical ontologies for Arabic. One critical obstacle facing Arabic ontology construction is the fact that there is no standard upper-level ontology to act as a foundation from which to build and this causes a severe lack of coherence and consistency among the Arabic ontologies. Most of the previous experiments in this field have structured the ontologies either fully manually or partially manually which leads to issues of high cost and a requirement for significant investment of time (Al-Zoghby, Elshiwi, and Atwan, 2018).

In the case of building an ontology in the Arabic language there is some previous work related to specific topics such as the Islamic domain, Computing domain, News domain, and Legal domain (Al-Zoghby, Elshiwi, and Atwan, 2018). Raisan and Abdullah (2017) implemented an ontology in the Arabic language related to the Iraqi News domain. They used a manual ontology development strategy to build the schema. A study by Albukhitan and Helmy (2013) produced an automatic ontology based annotation of Arabic Web resources related to food, nutrition and health domains. It used linguistic patterns to discover relevant relationships between the named entities in the Arabic Web resources.

Al-Safadi, Al-Badrani, and Al-Junidey (2011) presented a model for representing Arabic knowledge in the Computer Technology domain using ontologies. They stressed the importance of understanding the goal of the ontology and its end users during the development of the ontology. The study in (AlAgha and Abu-Taha, 2015) built an Arabic natural language interface to ontologies and RDF data (AR2SPARQL) by translating the user query to RDF triple patterns. The study applied Arabic NLP techniques to link query terms to ontological entities, then used a set of grammar rules in the ontology to structure a SPARQL query by joining the ontological entities.

On the other hand, there are many studies related to Latin-character-set-based language ontologies in multiple topic areas. Here we focus only on the previous work related to the infectious disease domain. Cowell and Smith (2010) reviewed the vocabulary resources relevant to the infectious disease domain and emphasised that they are lacking in terms of their support of translational medicine and computational applications. The study in (Schriml et al., 2011) generated a disease ontology including 8,043 human diseases with the goal of providing consistent, sustainable, and reusable descriptions of human disease terms.

The BioCaster project is a multilingual ontology that aims to describe the terms and relations necessary to detect at an early stage public health events in the grey literature, which is information or research that has been published in a non-commercial way, (Collier et al., 2010). It supports 12 languages including Arabic, but employs a translation-based method that may potentially produce incorrect words. In other words, some diseases produce another meaning after translation due to the complexity of the disease name. For example, the disease (اَلتهَاب السَحَايَا) which means Meningitis may be translated to Schizophrenia which is not correct. Moreover, using the abbreviations in English affect the results such as MARSA, which is Methicillin resistant Staphylococcus aureus, cannot be translated directly with out the full meaning.

Critically, none of the above studies have developed an infectious disease ontology in the Arabic language. Moreover, the existing infectious disease ontologies in other languages were constructed with only formal language input. Therefore, the common problems include concepts having different surface forms between formal and informal variants, or professional and non-professional descriptions of disease terminology and symptoms, and Arabic dialect detection in non-standard settings.

5.3 Understanding the Needs of an Infectious Disease Ontology in Arabic

The Arabic Infectious Disease ontology is classified as a domain ontology, wherein the concepts and relationships belong to a specific domain, in our case infectious disease. In other words, it decreases the potential confusion of terms between users who share different kinds of information that relate to this domain (Al-Zoghby, Elshiwi, and Atwan, 2018). Recently, ontologies have proved to be an effective approach to support data annotation to process data and information automatically. BioCaster is a notable example for using ontologies in detecting and tracking infectious disease (Cowell and Smith, 2010). Moreover, it has many advantages in data analysis of automated reasoning applications and high-throughput technologies such as clinical decision support, biomedical research, biosurveillance applications, and analysing microarray data tools. These benefits led to increasing attention on developing vocabularies and reasoning algorithms that are used in ontologies (Cowell and Smith, 2010). As a result, we have conducted experiments to understand what needs to be done to improve the strictness and the structure of ontologies that are effective in supporting the computational reasoning, analysis and monitoring of infectious disease.

The large scale survey conducted by Joshi, Karimi, Sparks, Paris, and C. Raina Macintyre (2019b) suggested that there is need to improve the quality of the epidemic intelligence by the mitigation of false information. The survey also highlighted that it is important to include informal text such as social media in the medical ontologies which contain scientific and technical terms. Slang terms, which are the informal language used by the general public, pervade the language of social media since there are lower expectations of formality in such writing and the expressions of emotions are more frequent (Soliman et al., 2014). Using slang terms especially in the medical domain has the potential to cause health issues when a misunderstanding occurs. For example, the word (سيلَان) may be understood as mucus or blood coming from the nose and this represents symptoms from multiple different diseases. The specific aims for our Arabic Infectious Disease Ontology are:

- To describe the terms and relations necessary to detect and analyse infectious disease.
- To bridge the gap between informal terms and scientific terms related to infectious disease.
- To be released in a form that is freely available.

5.4 Ontology Engineering

The process of ontology engineering includes many phases. It starts with understanding the domain knowledge to be represented by the ontology. The presentation of domain knowledge is represented by a basic classification of terminology in the domain and the relationships between the terms.

5.4.1 Ontology Architecture

The proposed architecture for the creation of our ontology is based on a hybrid approach, which merges two techniques, manual ontology development with ontology learning. Figure 5.1 represents the proposed system architecture for the creation of an Arabic Infectious Disease Ontology.

5.4.2 Data Source Selecting and Gathering

First, we defined a list of infectious diseases that typically spread in Arab countries based on information from the World Health Organization website[1]. We have developed a web crawler that retrieves relevant documents from trusted websites, that represent Governments, Organisations, or Hospitals, on Arabic infectious diseases listed in Table 5.1 in order to generate a high quality corpus. We have 115 documents that covered around 25 infectious diseases.

English Name	**Arabic Name**	**Link**
World Health Organization	منظمة الصحة العَالمية	https://www.who.int
The Ministry of Health in Saudi Arabia	وزَارة الصحة السعودية	https://www.moh.gov.sa
Wikipedia	ويكيبيديَا	https://ar.wikipedia.org
Mayo clinic	مَايو كلينك	https://www.mayoclinic.org/ar
WebTeb	ويب طب	https://www.webteb.com
Altibbi	الطبي	https://www.altibbi.com

Table 5.1: Some Arabic Web Sources Related to Infectious Diseases

[1] `https://www.who.int`

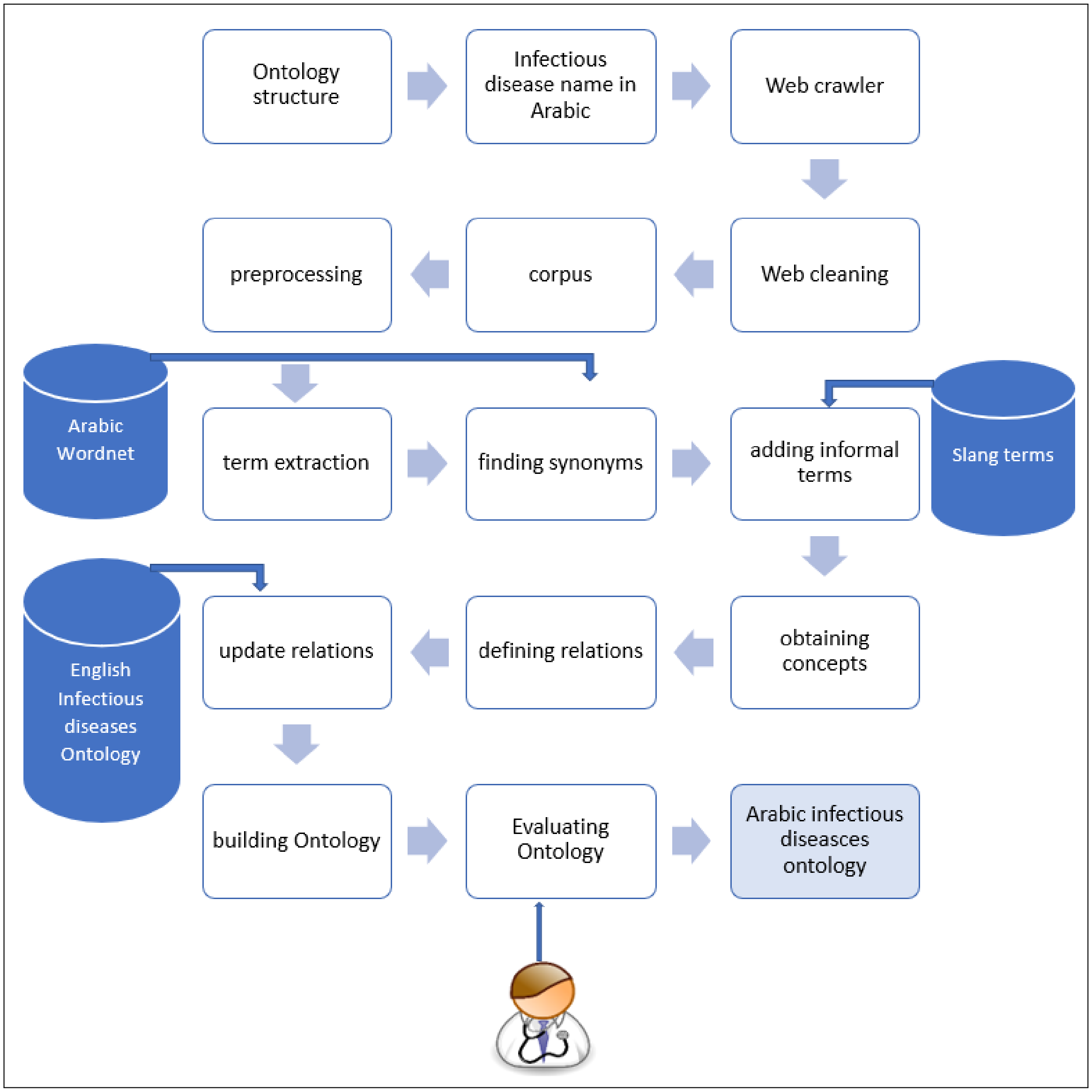


Figure 5.1: The proposed system architecture of Arabic Infectious Disease Ontology

5.4.3 Pre-Processing

We used standard approaches for pre-processing the text, applying the following processes using Python scripts[2] in sequence: Tokenization, Normalization, Stopword removal, POS Tagging, and Sentence Splitting in order to develop a high quality

[2] https://github.com/alsudias

ontology as described in (Asim et al., 2018). The pre-processing steps are as follows:

Tokenization: Tokenization is the process of chunking the text into words. This is a vital first step as we will be working with domain terms.

Normalization: In normalisation, some Arabic letters are normalised such as (أ ، آ ، إ) are converted to (ا).

Removing Stopwords: Arabic stopwords were removed since they are less useful for ontology creation. There is also need to remove the non-Arabic words, special characters, and numbers.

POS Tagging: Part of speech tagging is the process of labeling corpus words with their corresponding part of speech tags. We used the Stanford Arabic Part of speech tagger in this step (C. Manning et al., 2014).

Sentence Splitting: To allow better extraction and linking of terms, we used (.) to split the sentences in our corpus.

5.4.4 Term Extraction

Terms are linguistic representations of domain-specific concepts. The target here is to discover a set of significant terms for concepts and relations (Cimiano, 2006). We followed three statistical strategies for this step: TF-IDF, C-value and YAKE.

5.4.4.1 TF-IDF

The well known formula Term Frequency-Inverse Document Frequency (TF-IDF) is used to rank the candidate terms according to their importance in the corpus (Al-Thubaity et al., 2014). We apply unigrams, bigrams and trigrams and disregard terms that have a document frequency lower than two. Python's scikit-learn libraries 0.20.2 (Pedregosa et al., 2011) are used to apply TF-IDF on the data set.

5.4.4.2 Using C-value

C-value is a domain-independent method for automatic term recognition that aims to improve the extraction of nested multi-word terms (Frantzi, Ananiadou, and Mima, 2000). The algorithm first tags the words with their POS tags by Stanford Arabic Part of speech tagging (C. Manning et al., 2014). Then it extracts the strings using the linguistic filter, which is the type of terms extracted, and the stop words list. Then sequences that fall below the frequency threshold are removed. After that, we calculate C-value for each of the candidate strings. Finally, a list of candidate terms is built and ranked by their C-value.

5.4.4.3 Using YAKE

Yet Another Keyword Extractor (YAKE) is an Unsupervised Approach for Automatic Keyword Extraction using text features (Campos et al., 2019). YAKE has different steps including: pre-processing the text as before, and identifying the candidate term, feature extraction, computing term score, generating n-gram and computing candidate keyword score, and deduplicating and ranking the data. The ranking depends on the score for each term which is calculated from a defined set of features. These features are the casing aspect of a word, word position, word frequency, word relatedness to context, and the counts of word occurrences in different sentences. It is available online as an open source Python package. It includes support to extract keywords from Arabic language.

5.4.5 Finding Synonyms

Synonyms can be used to expand the terms that were not retrieved by web resources. We used Arabic WordNet (Black et al., 2006) to find synonyms of the term that will be used. For instance, the word (مرَاهم) , ointments in English, has a synonym (دهَان) and the two terms need to be added to ontology because they represent a treatment of Leishmania disease.

5.4.6 Adding Informal Terms

In order to find the slang terms related to infectious diseases, we consulted native speakers. This included eliciting the informal term equivalents of the infectious disease that may be used by the study participants when they write about disease in social media. A questionnaire was sent to five different types of people: academic, health professional, students, non-educational people and others as classified in the previous chapter (Chapter 4). The questionnaire template of this study can be found Appendix B. Around 100 people responded to the questionnaire with three informal names of each disease on average. For example, the disease (حمَى الضَنَك), Dengue fever in English, has six informal names: (الدَنجية), (أبو الركب), (الدنج), (الدنك), (حمَى تكسير العظَام), and (حمَى عدن).

In addition, we extend our approach by consulting these groups of people about the informal terms for the symptoms and then update the ontology with these terms. Words for symptoms are affected by the different dialects of the Arabic language since they depend largely on feelings and experiences. For instance, a person can discuss coryza using different informal words such as (زكَام), (رشح), and (عطَاس).

5.4.7 Obtaining Concepts

In order to decide which terms to be considered as a concept and included in the output ontology, we manually filter the terms that are meaningful. Table 5.2 shows the basic concepts of the proposed ontology. The ontology focuses on the infectious disease concept (المرض المعدي) due to the fact that it represents a central node in the ontology.

Concept	**Concept (in Arabic)**	**Description**
Infectious disease	المرض المعدي	Any disease that classified as infectious disease.
Slang term of Infectious disease	المصطلح العَامي لـلمرض المعدي	The informal or slang term of the infectious disease.
Symptom	عرض	Any symptom of the infectious disease.
Cause	سبـب	Anything that cause the infectious disease.
Prevention	وقاية	The ways of prevention of the infectious disease.
Infection	عدوَى	The ways that the infectious disease can infection between people.
Organ	عضو	The affected organ of the infectious disease.
Treatment	علاج	The methods taken to treat the infectious disease.
Diagnosis	تشخيص	The ways that help to define the infectious disease.
Place of the disease spread	مكان انتشار المرض	The place that the infectious disease may spread.
Infected category	الفِئة المصابة	The most vulnerable category that may infected.

Table 5.2: Arabic Infectious Disease Ontology Concepts

5.4.8 Defining and Updating Relations

This step is a manual one, we defined appropriate relations for the infectious disease domain. Some of these relations are derived from the ontologies in (AlAgha and Abu-Taha, 2015) and (Schriml et al., 2011). The other relations are defined after studying the goals of the ontology. For instance, the relations spread-in (ينتشر في) and infested-with (موبوءه بـ) may help in finding the regions where the infectious disease typically spread. Also, the relation synonym (يعرف ، تعرف) is important to find all the slang names of the infectious disease. Table 5.3 represents the relations with Arabic examples and description of the relation.

In order to expand our ontology with all terms and relations, we reviewed an English infectious disease ontology. For this, we used (Schriml et al., 2011) as infectious disease ontology and translated the concepts and relations to Arabic. Then, we updated the ontology with these components.

5.5 Ontology Storage and Representation

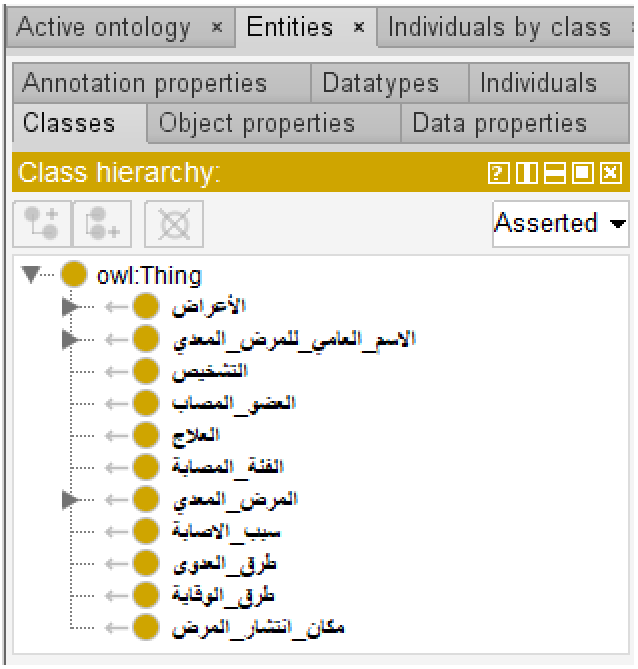
We stored the ontology in the Web Ontology Language (OWL) format in order to support sharing and reuse of this valuable resource. We used the Protégé tool [3] version 5.5.0, a free open-source software tool typically used for building and maintaining ontologies. It supports many languages including Arabic. Protégé has particular benefits over other tools because it facilitates shareable access to structured knowledge for domain specialists, software designers, and knowledge engineers simultaneously (Kapoor and Sharma, 2010). Also, it is capable of supporting XML, RDF (S), XML Schema and OWL formats (Musen et al., 2015).

The classes have been organised in a hierarchical taxonomy as in Figure 5.2. The properties in OWL can be divided into Object property and Datatype property. As a result, in this study there are 21 Object properties that have been proposed in Table 5.3 and 11 Datatype properties that have been proposed as illustrated in Figure 5.3. The Object properties define the relations between classes in order to give a sufficient vision of how the components of the ontology interact with each other. There is a need to take into account the inverse characteristic for each property. For instance, 'cause' and 'caused-by' are two properties that work in opposite directions between the 'infectious disease' and 'cause' classes. In addition to the Object property, the Datatype property has been proposed to illustrate the type of data for individuals. In our case, the range for all Datatype property is a string.

[3] https://protege.stanford.edu/

Relation	**Relation (in Arabic)**	**Examples**	**Links Description**	
			From	To
Is-a	هو	هو ، هي	infectious disease	definition (hyponym relationship)
Kind-of	نوع من	هو من ، هي من ، هو نوع من ، هي نوع من	infectious disease	type of the disease.
Synonym	مرادف	يعرف ، تعرف	infectious disease	informal term (slang one)
Caused-by	تحدث بسبب	تحدث بسبب، يحدث بسبب	infectious disease	cause class
Cause	يسبب	يسبب ، تسبب	cause class	infectious disease
Infected-by	يُعـدَى بـ	يُعـدَى بـ ، تُعـدَى بـ	infectious disease	infection class
Infect	يعدي	يعدي ، تعدي ، طرق العدوَى	infection class	infectious disease
Prevent	يقي ، يمنع	يقي ، تقي	prevention class	infectious disease
Prevented-by	يقي بـ	يقي بـ ، يقي بـ ، طرق الوقَاية	infectious disease	prevention
Spread-In	ينتشر في	ينتشر في، تنتشر في	infectious disease	the place
Infested-with	موبوءه بـ	موبوءه بـ ، أَمَاكن انتشَار	the place	the infectious disease
Symptom-of	عرض لـ	أَحدَى أَعرَاض ، أَحدَى أَعرَاضهَا	Symptom class	infectious disease
Has-symptom	له عرض	من أَعرَاضه ،من أَعرَاضهَا	infectious disease	Symptom class
Diagnoses	يشخص	طرق تشخيص	the Diagnosis	infectious disease
Diagnoses-by	يشخص بـ	لتشخيص	infectious disease	the Diagnosis
Infects	يصيب	يصيب ، تصيب ، يمرض ، تمرض	infectious disease	the organ
Infected-by	يصَاب بـ	يصَاب بـ ، تصَاب بـ	the organ	infectious disease
Treats	يعَالج	يعَالج ، تعَالج ، يدَاوي ، تدَاوي	treatment class	infectious disease
Treated-by	يعَالج بـ	يعَالج بـ ، تعَالج بـ طرق العلَاج ،	infectious disease	treatment class
Get-sick	تصيب بَالمرض	تصيب بَالمرض ، يصيب بَالمرض	infectious disease	Infected category
Get-sick-by	تصَاب بَالمرض	تصَاب بَالمرض ، يصَاب بَالمرض	Infected category	infectious disease

Table 5.3: Arabic infectious disease Ontology relations



- owl:Thing
  - الأعراض
  - الاسم_العامي_للمرض_المعدي
  - التشخيص
  - العضو_المصاب
  - العلاج
  - الفئة_المصابة
  - المرض_المعدي
  - سبب_الاصابة
  - طرق_العدوى
  - طرق_الوقاية
  - مكان_انتشار_المرض

Figure 5.2: Arabic Infectious Disease Ontology Classes in Hierarchy

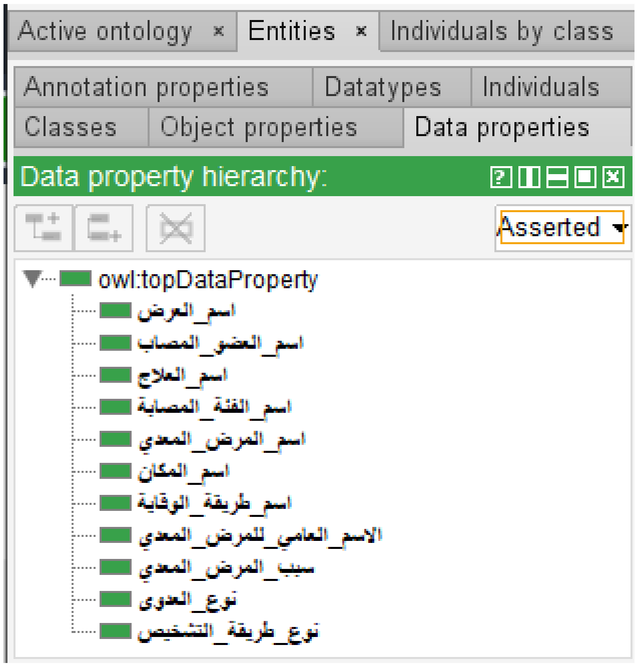
Figure B.1 shows a sub graph of the developed ontology with classes and their relationships, and Figure B.2 illustrates a screen shot of the Measles disease (الحَصبَة) as an example.

5.6 Ontology Evaluation

Many techniques were used to evaluate ontologies. Here we conducted performance of the term extraction technique and a Domain Expert Evaluation approach.

5.6.1 Evaluating the Term Extraction Results

In order to evaluate the performance of the term extraction techniques, the terms in the extracted ontology are manually reviewed. Then, we compared the extracted terms by the three strategies, TF-IDF, C -value and YAKE, to the final concepts using the standard Precision, Recall, and F-Measure metrics. Table 5.4 summarises the results obtained by the system and Figure 5.4 visualizes the average results of the term extraction performance.



owl:topDataProperty
- اسم_العرض
- اسم_العضو_المصاب
- اسم_العلاج
- اسم_الفئة_المصابة
- اسم_المرض_المعدي
- اسم_المكان
- اسم_طريقة_الوقاية
- الاسم_العامي_للمرض_المعدي
- سبب_المرض_المعدي
- نوع_العدوى
- نوع_طريقة_التشخيص

Figure 5.3: Arabic Infectious Disease Ontology Datatype properties (as translated above).

We can see that the lowest performance was found in the C-value method. This could result from the complexity of the information of infectious disease in Arabic language and the document size. The TF-IDF method achieved between 20.0% to 50.0% in their results due to the different types of documents included. In the YAKE method, the overall performance for the Precision is around 35.0% while the Recall is achieved up to 62.9% in Mumps documents. Moreover, the YAKE method supports multiword term extraction which is important in Arabic since many terms in infectious disease consisting of two or three words. For example, the word (التهَاب الملتحمة), which means conjunctivitis in English, is a multi word that describes one symptom of the Zika virus disease.

5.6.2 Domain Expert Evaluation

We asked an expert in infectious diseases who is an Arabic native speaker to evaluate our ontology in terms of medical correctness. The evaluation included reviewing the concepts and the relations between the diseases. In other words, the expert needed to review every infectious disease and terms related to it and the correct connection between them. Also, we wished to check the accuracy of our ontology in terms of the

Disease Name	**TF-IDF**			**C-Value**			**YAKE**		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
AIDS	32.0%	23.3%	26.9%	18.0%	12.4%	14.7%	40.0%	28.5%	33.3%
Avian influenza	21.0%	16.9%	18.7%	12.5%	12.8%	12.7%	28.0%	32.0%	29.9%
Coronavirus	25.0%	28.0%	26.4%	15.0%	18.6%	16.6%	32.0%	39.2%	35.2%
Dengue fever	30.0%	28.1%	29.0%	13.7%	23.2%	17.3%	34.0%	29.3%	31.5%
Elephant disease	36.0%	27.2%	30.9%	09.0%	06.0%	07.7%	33.0%	23.9%	27.8%
German measles	30.0%	33.6%	31.7%	18.0%	21.5%	19.6%	36.0%	42.6%	39.0%
HPV infection	30.0%	29.7%	29.8%	16.3%	15.0%	15.6%	27.0%	26.4%	26.7%
Leishmania	32.0%	32.4%	32.2%	21.0%	20.8 %	20.9%	38.0%	40.2%	39.1%
Lice	22.0%	23.9%	22.9%	13.0%	15.2%	14.0%	25.0%	28.4%	26.6%
Malaria	40.0%	53.1%	45.6%	21.2%	35.9%	26.7%	38.0%	51.4%	43.7%
Marsa	28.0%	29.6%	28.8%	12.0%	15.7%	13.6%	32.0%	38.8%	35.1%
Maltese fever	32.0%	45.2%	37.5%	09.0%	09.7%	09.3%	37.0%	50.7%	42.8%
Measles	43.0%	41.5%	42.3%	29.0%	27.6%	28.3%	46.0%	42.3%	44.0%
Meningitis	35.0%	35.3%	35.1%	16.0%	15.5%	15.7%	40.0%	34.5%	37.0%
Mumps	31.0%	49.0%	37.9%	16.3%	25.4%	19.8%	39.0%	62.9%	48.2%
Nipah virus	26.0%	40.4%	31.6%	15.0%	31.3%	20.3%	26.0%	44.4%	32.8%
Plague	28.0%	33.5%	30.5%	16.0%	17.8%	16.8%	34.0%	37.2%	35.5%
Rabies	31.0%	32.8%	31.9%	23.0%	24.6%	23.8%	33.0%	34.7%	33.8%
Schistosomiasis	22.0%	32.3%	26.2%	06.0%	08.4%	06.9%	33.0%	50.3%	39.9%
Seasonal flu	50.0%	52.6%	51.3%	11.7%	13.1%	12.3%	43.0%	47.4%	45.1%
Tuberculosis	34.0%	36.5%	35.2%	15.0%	13.9%	14.4%	32.0%	30.3%	31.1%
Yellow fever	25.0%	24.9%	24.9%	11.2%	14.1%	12.5%	34.0%	34.3%	34.2%
Zika virus	27.0%	23.3%	25.0%	20.0%	16.3%	17.9%	34.0%	28.7%	31.1%

Table 5.4: Term Extraction Performance

individual infectious diseases. For example, is it correct that Malaria is infected by the Transfusion.

In addition, we calculated the precision of the system after computing the average number of correct terms which was 88.4%. This result shows that the ontology has a high correctness since the expert agreed on the applicability of using it.

5.7 Conclusion

In this chapter, we have described a novel resource that integrates the scientific vocabularies of infectious diseases with their informal equivalents. The Arabic Infectious Disease Ontology is written in the Arabic language, and is the first ontology in Arabic specialising in the infectious disease domain. It contains 11 classes, 21 object properties, 11 datatype properties, and 215 individual concepts. The ontology is made freely available for academic research [4].

[4] https://doi.org/10.17635/lancaster/researchdata/350

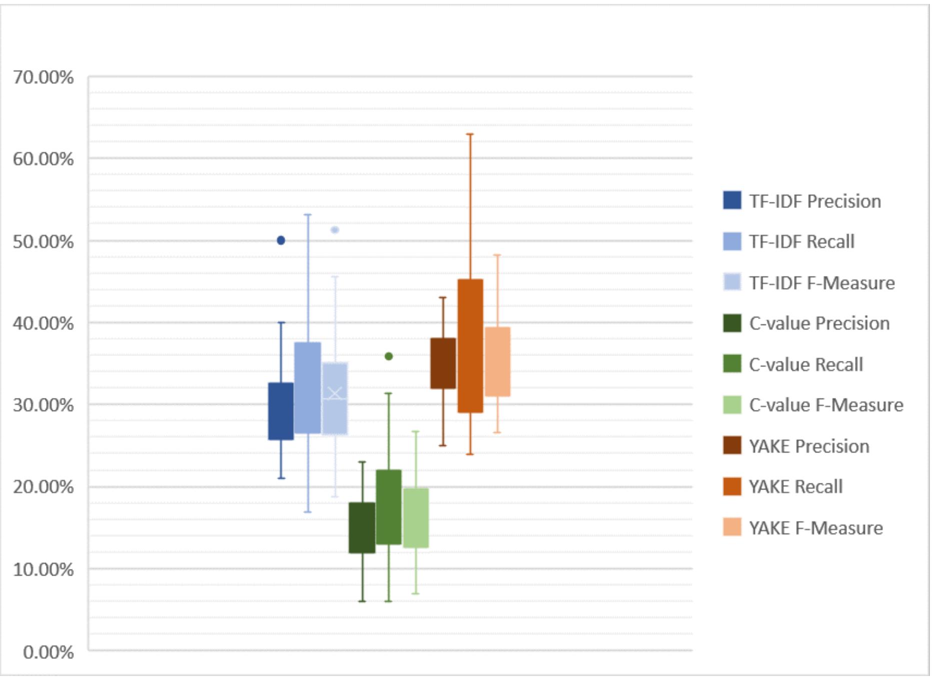


Figure 5.4: Average of Term Extraction Performance of Infectious diseases

The combination of ontology learning strategies and manual techniques have enhanced the accuracy of our results. Our focus in this chapter is to implement the Arabic Infectious Disease Ontology to support the analysis of the spread of infectious disease in any Arab country. Therefore, it is important that an extended list of terms that are related to each infectious disease is used, since the combination of the informal names of the diseases with the formal ones will help in expanding the retrieval of information during data collection.

The ontology is flexible and easy to update when needed. For example, with the recent appearance of COVID-19 as a new disease, there is need to update the Arabic Infectious Disease Ontology with COVID-19 information. This included symptom, cause, prevention, infection, organ, treatment, diagnosis, place of the disease spread, and slang terms for COVID-19. A detailed explanation of this step is clarified in Section 6.3

In the next chapters, Chapter 6 and Chapter 7, there are some experiments using the Arabic Infectious Disease Ontology in various applications. In Chapter 6, we analyse pandemic tweets in three different ways: identifying the topics discussed during the period, detecting rumours, and predicting the source of the tweets.

Whilst in Chapter 7, we analyse Arabic tweets to estimate their usefulness for health surveillance and understand the impact of the informal terms in the analysis.

Chapter 6

Analysing Tweets During a Pandemic

6.1 Introduction

This chapter is about analysing COVID-19 Arabic tweets that have been written during the pandemic. The goal of this chapter is to combine qualitative and quantitative studies to analyse Arabic tweets with the aim of supporting Public Health Organisations (PHOs) who can learn from social media data. It acts as a case study to demonstrate and evaluate how we can bring together the techniques and resources presented in the previous two chapters. This chapter is derived from the published paper under the title "COVID-19 and Arabic Twitter: How can Arab World Governments and Public Health Organizations Learn from Social Media?", the paper was published at the NLP COVID-19 Workshop, an emergency workshop at ACL 2020[1] (Alsudias and Rayson, 2020b).

In March 2020, the World Health Organization announced the COVID-19 outbreak as a pandemic. Governments around the world have taken different decisions in order to stop the spread of the disease. Many academic researchers in a variety of fields, including NLP, have conducted research on this subject.

People use social media applications such as Twitter to find the news related to COVID-19 and/or express their opinions and feelings about it. As a result, a vast amount of information could be exploited by NLP researchers for a myriad of analyses despite the informal nature of social media writing style.

[1] https://www.nlpcovid19workshop.org/

COVID-19 Twitter conversations may not correlate with the actual disease epidemiology. Therefore, PHOs have a vested interest in ensuring that information spread in the population is accurate (Vorovchenko et al., 2017). For instance, the Ministry of Health in Saudi Arabia[2] presented a daily press conference incorporating the aim of quickly stopping the spread of false rumours. However, there was a prolonged period of time until these warnings are issued. For example, the first tweet that included false information about hot weather killing the virus was on 10 February 2020, while the press conference which responded to this rumour on 14 April 2020. There is a clearly a need to find false information as quickly as possible. In addition, effort needs to be made in relation to tracking the user accounts that promote rumours. This can be undertaken using a variety of techniques, for example using social network features, geolocation, bot detection or content based approaches such as language style. PHOs would benefit from speeding up the process of tracking in order to stop rumours and remove the bot networks.

The vast majority of the previous research in this area has been on English Twitter content but this will not directly assist PHOs in Arabic speaking countries. The Arabic language is spoken by 467 million people in the world and has more than 26 dialects, as explained in Section 2.2.1. Of particular importance for NLP is coping with dialectal and/or meaning differences in less formal settings such as social media. As an example in the health field, the word (تحصين) may be understood as vaccination[3] in modern standard Arabic or reading supplications in Najdi dialect[4]. There has been much recent progress in Arabic NLP research, yet there is still an urgent need to develop fake news detection for Arabic tweets (Mouty and Gazdar, 2018).

We hypothesise that PHOs may benefit from mining the topics discussed between people during the pandemic. This may help in understanding a population's current and changing concerns related to the disease and help to find the best solutions to protect people. In addition, during an outbreak, people themselves will search online for reliable and trusted information related to the disease such as prevention and transmission pathways. Therefore, in order to help PHOs, we have analysed data from social media along various lines: Analysing the topics discussed between people during the peak of COVID-19, identifying and detecting the rumours related to COVID-19, and predicting the type of sources of tweets about COVID-19.

[2] https://www.moh.gov.sa

[3] https://en.wikipedia.org/wiki/Modern_Standard_Arabic

[4] https://en.wikipedia.org/wiki/Najdi_Arabic

6.2 Related Work

There is a vast quantity of research over recent years that analyses social media data related to different pandemics such as H1N1 (2009), Ebola (2014), Zika Fever (2015), and Yellow Fever (2016). These studies followed a variety of directions for analysis with multiple different goals (Joshi, Karimi, Sparks, Paris, and C Raina Macintyre, 2019a). The study of (Ahmed, P. A. Bath, Sbaffi, et al., 2019) used a thematic analysis of tweets related to the H1N1 pandemic. Eight key themes emerged from the analysis: emotion, health related information, general commentary and resources, media and health organisations, politics, country of origin, food, and humour and/or sarcasm.

A survey study (I. C.-H. Fung et al., 2016) reviewed the research relevant to the Ebola virus and social media. It compared research questions, study designs, data collection methods, and analytic methods. Ahmed, G. Demartini, and P. Bath (2017) used content analysis to identify the topics discussed on Twitter at the beginning of the 2014 Ebola epidemic in the United States. In (Vorovchenko et al., 2017), they determined the geolocation of the Ebola tweets and named the accounts that interacted more on Twitter related to the 2014 West African Ebola outbreak. The main goal of the study by Kalyanam et al. (2015) was to distinguish between credible and fake tweets. It highlighted the problems of manual labeling process with verification needs. The study in (I.-H. Fung et al., 2016) highlighted how the problem of misinformation changed during the disease outbreak and recommended a longitudinal study of information published on social media. Moreover, it pointed out the importance of understanding the source of this information and the process of spreading rumours in order to reduce their impact in the future.

Ghenai and Mejova (2017) tracked Zika Fever misinformation on social media by comparing them with rumours identified by the World Health Organization. Also, they pointed out the importance of credible information sources and encouraged collaboration between researchers and health organisations to rapidly process the misinformation related to health on social media. The study in (Ortiz-Martınez and Jiménez-Arcia, 2017) reviewed the quality of available yellow fever information on Twitter. It also showed the significance of the awareness of misleading information during pandemic spread. The study of (Zubiaga, Aker, et al., 2018a) summarised other studies related to social media rumours. It illustrated techniques for developing rumour detection, rumour tracking, rumour stance classification, and rumour veracity classification. Vorovchenko et al. (2017) mentioned the importance of Twitter information during the epidemic and how Public Health Organisations can benefit from this. They showed the requirement to monitor false information posted by some accounts and recommended that this was performed in real time to reduce the danger of this information. Also, they discussed the lack of available datasets which help in the development of rumour classification systems.

Researchers have been doing studies on building and analysing COVID-19 Twitter datasets since the disease appeared in December 2019. So far, there are two different datasets which have been published related to Arabic and COVID-19 (Alqurashi, Alhindi, and Alanazi, 2020) and (Haouari et al., 2020b). The former collected 3,934,610 tweets until April 15 2020, and the latter included around 748k tweets until March 31, 2020. These papers contain an initial analysis and statistical results for the collected tweets and some suggestions for future work, which include pandemic response, behaviour analysis, emergency management, misinformation detection, and social analytics.

In addition, there are some datasets in English such as (Chen, Lerman, and Ferrara, 2020) and (Lopez, Vasu, and Gallemore, 2020). Also, there is a multilingual COVID-19 dataset containing location information (Qazi, Imran, and Ofli, 2020). This contains more than 524 million tweets, with 5.5 million Arabic tweets, posted over a period of three months since February 1, 2020. It focuses on determining the geolocation of a tweet which can help research with various different challenges, including identifying rumours.

Although the above studies have produced datasets related to COVID-19, they do not analyse them deeply using NLP methods. Previous studies representing earlier epidemics present good techniques and results, however none of them are related to Arabic tweets. Therefore, to assist PHOs in Arabic speaking countries there is an urgent need to analyse tweets related to COVID-19 using multiple Arabic NLP techniques.

6.3 Updating the Arabic Infectious Disease Ontology

With the appearance of COVID-19 as a new disease, there is a requirement to update our Arabic Infectious Disease Ontology (Alsudias and Rayson, 2020c), which integrates the scientific and medical vocabularies of infectious diseases with their informal equivalents used in general discourse. We collated COVID-19 information from the World Health Organization[5] and Ministry of health in Saudi Arabia. This included symptom, cause, prevention, infection, organ, treatment, diagnosis, place of

[5] http://www.emro.who.int

the disease spread, and slang terms for COVID-19 and extended our ontology[6]. These terms were then used in our collection process.

6.4 Data Collection

We began collecting Arabic tweets about a number of infectious diseases from September 2019, as described in Chapter 3. Here in this chapter, we analysed only the tweets related to COVID-19 from December 2019 to April 2020 (there are a few tweets between September and November, these are related to Middle East respiratory syndrome coronavirus, MERS-CoV[7]). We have collected approximately six million tweets in Arabic during this period. We obtained the tweets depending on three keywords (كورونَا, كرونَا, كوفِيد ١٩) which mean Coronavirus, a misspelling of the name of Coronavirus, and COVID-19 respectively in English. We collected the tweets weekly using the Twitter API.

Next, we pre-processed the tweets through a pipeline of different steps:

- Manually remove retweets, advertisements, and spam.
- Filter out URLs, mentions, hashtags, numbers, emojis, repeating characters, and non-Arabic words using Python scripts.
- Normalize and tokenize tweets.
- Remove Arabic stopwords (Alrefaie, 2017).

After pre-processing, the resulting dataset was 1,048,575 unique tweets from the original 6,578,982 collected. Figure 6.1 shows the number of Arabic tweets about Coronavirus each week with specific dates highlighted to show government decisions on protecting the population from COVID-19 and other key dates for context.

6.5 Methods

We performed three different types of analysis on the collected data. Firstly, in order to better understand the topics discussed in the corpus, we carried out a cluster analysis. Secondly, taking a sample of the corpus, we performed rumour detection. Finally, we extended our previous work, discussed in Chapter 4, to classify the source of tweets into five types of Twitter users which aims at helping to determine their veracity.

[6] `https://github.com/alsudias/Arabic-Infectious-Disease-Ontology`

[7] `https://bit.ly/3xCc7sr`

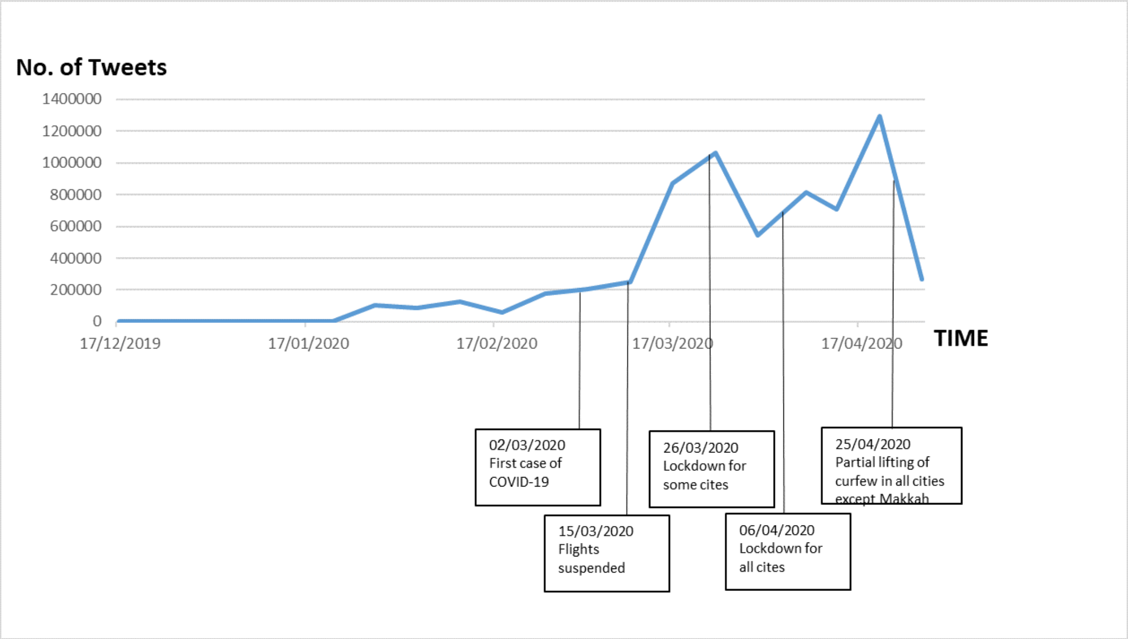


Figure 6.1: Number of Arabic tweets about Coronavirus

6.5.1 Cluster Analysis

To explore the topics discussed on Twitter during the COVID-19 epidemic in Saudi Arabia and other countries in the Arab World, we subjected the text of the tweets to cluster analysis. After pre-processing the tweets as described above, we used the N-gram forms (unigram, bigram, and trigram) of Twitter corpus and clustered them using the K-means algorithm with the Python Scikit-learn 0.20.2 (Pedregosa et al., 2011) software and set the value of k, the number of clusters, to be five. We selected this number after running the algorithm many times and analysing the results to avoid overlapping between clusters.

6.5.2 Rumour Detection

Following previous work (Zubiaga, Aker, et al., 2018b), we applied a top-down strategy, which is where the set of rumours is identified in advance then the data is sampled to extract the posts associated with the previously identified rumours. In our dataset, out of the one million tweets, we sampled 2,000 tweets to classify them for rumour detection. We manually labelled the tweets to create a gold standard dataset and then applied different machine learning algorithms in this part of our study.

6.5.2.1 Labelling Guidelines

We manually labelled the tweets with 1, -1 , and 0 to represent false information, correct information, and unrelated content, respectively. Our reference point for deciding whether the content contained true or false information was based on the list issued by the Ministry of Health in Saudi Arabia [8] and is regularly updated (the last update applied for this study dates from 14 April 2020). Table 6.1 presents the list in both Arabic and English and Table 6.2 shows some example tweets for each label.

Rumour in Arabic	**Rumour in English**
الحيوَانَات الاَليفة تنقل فيروس كورونَا.	Pets are transporters of Coronavirus.
البعوض نَاقل لّكورونَا.	Mosquitoes are transporters of Coronavirus.
الاَطفَال قد لَا يصَابون بفيروس كورونَا.	Children are not infected by Coronavirus.
كبَار السن فقط قد يتعرضون لمخَاطر. سيئة من فيروس كورونَا.	Only old people may have a high risk of Coronavirus.
حرَارة الطقس أو برودته تقضي علَى الفيروس.	Hot or cold weather can kill the virus.
الغرغرة بَالمَاء و الملح تقضي علَى الفيروس.	Gargling with water and salt eliminates the virus.
يوجد خلطَات و أعشَاب تقي من الكورونَا.	There are some herbs that protect against from Coronavirus.
الفيروس لَا يبقَى علَى الآسطح.	The virus does not survive on surfaces.

Table 6.1: List of rumours that appear during COVID-19 (source: Saudi Arabia Ministry of Health)

6.5.2.2 Machine Learning Models

We applied three different machine learning algorithms: Logistic Regression (LR), Support Vector Classification (SVC), and Naïve Bayes (NB). To help the classifier distinguish between the classes more accurately, we extracted further linguistic features. The selected features fall into two groups: word frequency (count vector

[8] https://www.moh.gov.sa

Tweet in Arabic	**Tweet in English**	**Label**
سيكون هنَاك انحسَار لَانتشَار فيروس كورونَا مع بدَاية فصل الصيف خصوصَا في العَالم العربي نظرَا لَارتفَاع درجة الحرَارة.	There will be a decrease in the spread of the Corona virus at the beginning of the summer, especially in the Arab world, due to the high temperatures.	1 (false information)
الصحة: يعيش الفيروس و يرتكز بَالَاسَاس في الجهَاز التنفسي لذلك غير وَارد انتقَاله عن طريق الحشرَات أو من خلَال لدغة البعوض.	The Ministry of Health: A virus lives and is mainly concentrated in the respiratory system, so it is not likely to be transmitted by insects or by mosquito bites.	-1 (true information)
اللّهم في هذي السَاعة المبَاركة نسآلك ان ترحمنَا و تبعد عنَا كل دَاء و بلَاء و قنَا شر الَامرَاض و الَاسقَام. وَاحفظ بلَادنَا و كَافة بلَاد المسلمين.	Oh God, in this blessed hour, We ask you to have mercy on us and keep away from us all disease and calamity, and protect us from the evil of diseases and sicknesses. Preserve our country and other Muslim countries.	0 (unrelated)

Table 6.2: Example tweets and our labelling system

and TF-IDF) and word embedding based (Word2Vec and FastText). We used 10-fold cross validation to determine accuracy of the classifiers for this dataset, splitting the entire sample into 90% training and 10% testing for each fold.

6.5.3 Source Type Prediction

We replicated a Logic Regression model from our study explained in Chapter 4, which was useful for classifying tweets into five categories: academic, media, government, health professional, and public (Alsudias and Rayson, 2019). We used the LR model because it previously achieved the best accuracy (77%), and employed it here to predict the source of the COVID-19 tweets that we had already labelled in Section 6.5.2.

6.6 Results and Discussion

6.6.1 Cluster Analysis

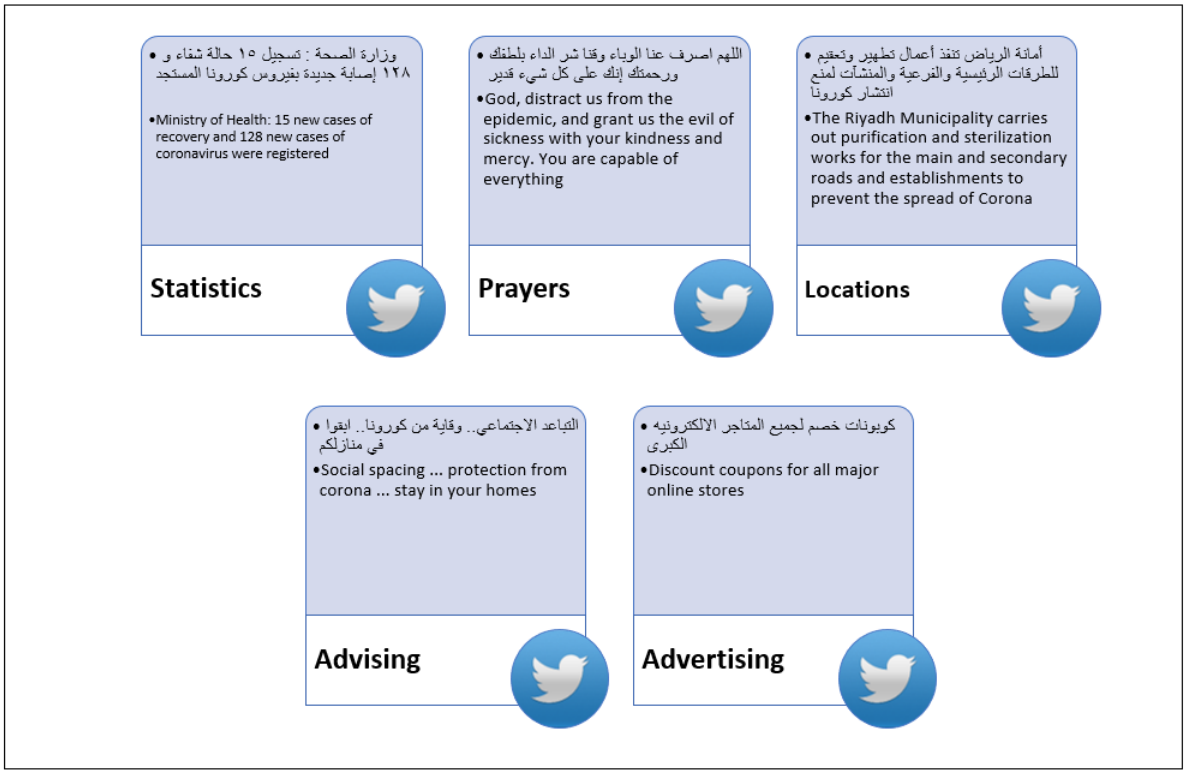
Our cluster analysis of the five main public topics discussed in tweet content is as follows: (1) disease statistics: the number of infected, died, and recovered people; (2) prayers: prayer asking God to stop virus; (3) disease locations: spread and location (i.e., name of locations, information about spread); (4) Advice for prevention education: health information (i.e., prevention methods, signs, symptoms); and (5) advertising: adverts for any product either related or not related to the virus. Figure 6.2 illustrates the five topics of COVID-19 tweets with examples.

For each cluster, the top terms by frequency are as follows: (1) disease statistics: case (حَالة), new (جديدة), and infection (اصَابة); (2) prayers: Allah (الله), Oh God (اللّهم), and Muslims (المسلمين); (3) disease locations: Dammam (الدمَام), Riyadh (الريَاض), and Makkah (مكة); (4) Advise for prevention education: crisis (ازمة), spread (انتشَار), and pandemic (جَاِئحة); (5) advertising: discount (خصم), coupon (كوبون), and code (كود).

We found that four of our categories (disease statistics, prayers, disease locations, and advice for prevention education) are similar to those found by Odlum and Yoon (2015) which are risk factors, prevention education, disease trends, and compassion. The marketing category is one of the topics in (Ahmed, Demaerini, and P. A. Bath, 2017) which discussed the topics in Twitter during the Ebola epidemic in the United States. Jokes and/or sarcasm is one of the categories that did not appear in our study but can be found in (Ahmed, Demaerini, and P. A. Bath, 2017) and (Ahmed, P. A. Bath, Sbaffi, et al., 2019), a thematic analysis study of Twitter data during H1N1 pandemic. This may be a result of more concern and panic from COVID-19 than other diseases during this period of time.

6.6.2 Rumour Detection

The result of our manual labelling process is 316 tweets label with 1 (false), 895 tweets label with -1 (true), and 789 tweets label with 0 (unrelated). Therefore,



- وزارة الصحة : تسجيل ١٥ حالة شفاء و ١٢٨ إصابة جديدة بفيروس كورونا المستجد
- Ministry of Health: 15 new cases of recovery and 128 new cases of coronavirus were registered

Statistics

- اللهم اصرف عنا الوباء وقنا شر الداء بلطفك ورحمتك إنك على كل شيء قدير
- God, distract us from the epidemic, and grant us the evil of sickness with your kindness and mercy. You are capable of everything

Prayers

- أمانة الرياض تنفذ أعمال تطهير وتعقيم للطرقات الرئيسية والفرعية والمنشآت لمنع انتشار كورونا
- The Riyadh Municipality carries out purification and sterilization works for the main and secondary roads and establishments to prevent the spread of Corona

Locations

- التباعد الاجتماعي.. وقاية من كورونا.. ابقوا في منازلكم
- Social spacing ... protection from corona ... stay in your homes

Advising

- كوبونات خصم لجميع المتاجر الالكترونيه الكبرى
- Discount coupons for all major online stores

Advertising

Figure 6.2: Examples of tweets in each cluster

the false information represents about 15.8% (from 2,000 tweets) and around 26% (from 1,211 tweets, after removing the unrelated ones). In the study by Ortiz-Martınez and Jiménez-Arcia (2017), 61.3% (from 377 tweets) of data was classified as misinformation about Yellow Fever. It represented 32% (from 26,728 tweets) considered as rumours related to Zika Fever in (Ghenai and Mejova, 2017).

Figures 6.3, 6.4, and 6.5 show the accuracy, F1-score, recall, and precision on our corpus using LR, SVC, and NB algorithms with various feature selection approaches. The highest accuracy (84.03%) was achieved by the LR classifier with a count vector set of features and SVC with TF-IDF. Therefore, the count vector gives best result in LG and NB in all metrics results whereas with precision which achieves better results with TF-IDF 83.71% in LG and 81.28% in NB. while TF-IDF in SVC has the best results except recall which achieved 75.55% in count vector set features.

We also applied several word embedding based approaches but without obtaining good results. The accuracy ranges from 48.90% to 60.97% and the F1 score is around

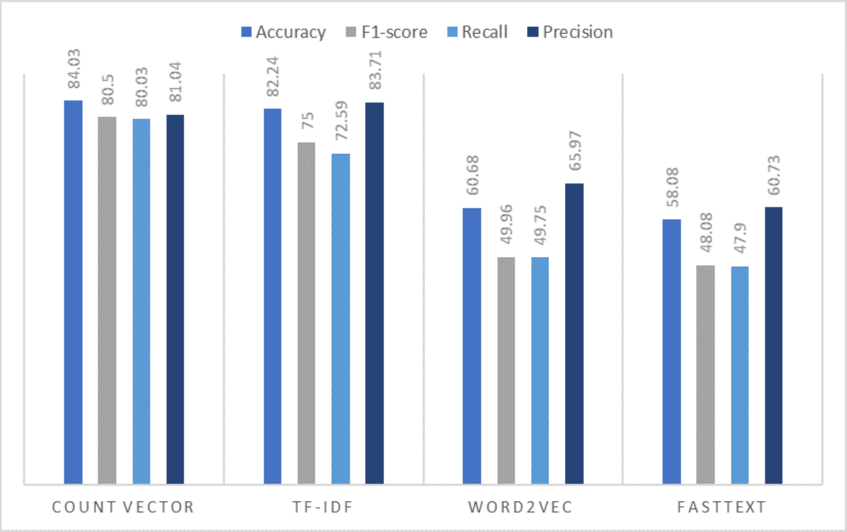


Figure 6.3: Results using Logistic Regression

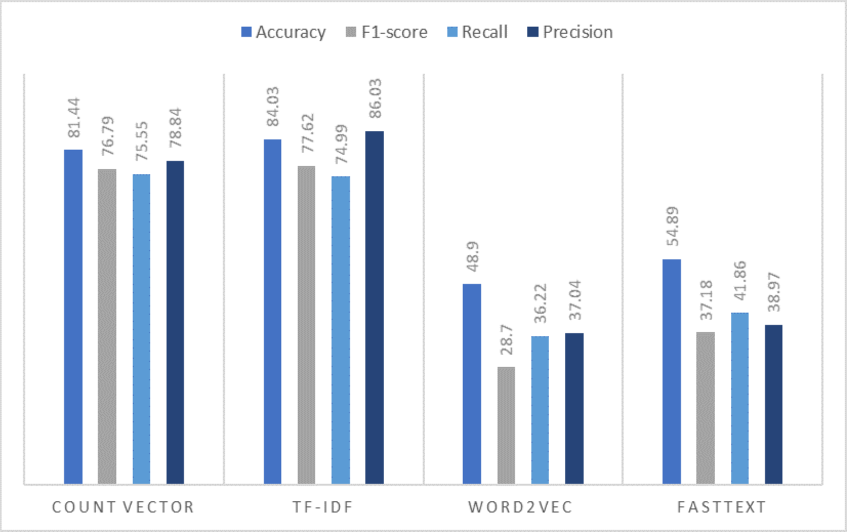


Figure 6.4: Results using Support Vector Classification

40.50% on average. FastText models achieve better accuracy in SVC (54.89%) and NB (59.49%) than Word2Vec by approximately (5%). While Word2Vec shows the best result with LG 60.68% for accuracy, 49.85% for F1 and recall, and 65.97% in precision.

Word embeddings are constructed based on the co-occurrence of words in a fixed window. For instance, opposing terms may appear together or close together. Therefore, the vector for the word "good" can be close to the vector for the word

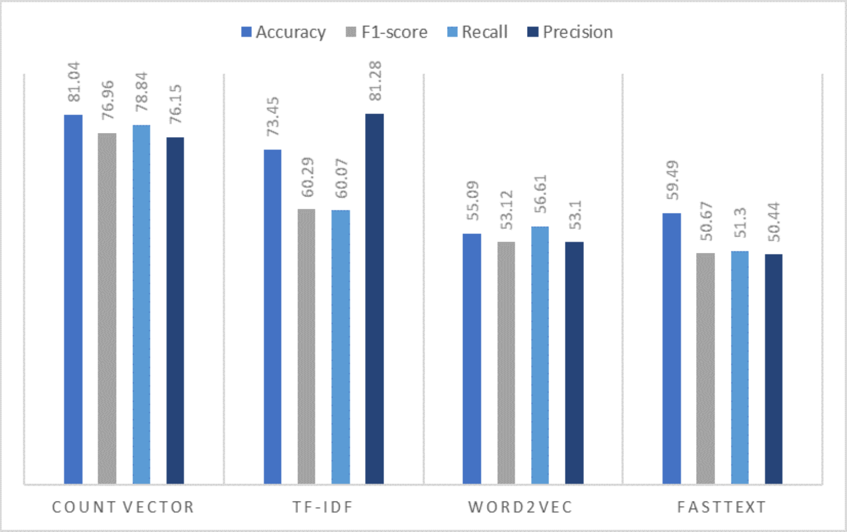


Figure 6.5: Results using Naïve Bayes

“bad”. This is problematic for applications such as sentiment analysis, in which such words correspond to orthogonal classes. As a result, word embeddings are not always a good fit for applications that need the use of antonyms or even synonyms (Uprety, Gkoumas, and Song, 2020).

The word frequency based approaches have around a 20% better result than the word embedding ones. The reason for this is expected to be the dataset size and the specific domain of context (Ma, 2018). We assumed that the word embedding methods may achieve good results due to the importance of the relevant information around the word. For example, FastText can deal with the misspelling problem which is common in social media language style and improves word vectors with subword information (Bojanowski et al., 2017).

6.6.3 Source Type Prediction

The model predicts the source type for each of the tweets. Table 6.3 shows some examples of the tweets with predicted labels by the model. We focus on the result of the fake news content since they are of highest importance for the Public Health Organisations. 30% (95 of 316) and 28% (91 of 316) of the rumour tweets are classified as written by a health professional and academic consequently. While only 12% (39 of 316) of them are predicted as written by the public. With this result, we find that the tweets containing false information quite often used the language style of academics and health professionals.

Tweet in Arabic	**Tweet in English**	**Predicted Label**
في قرَاءة علمية… يتوقع اندثَار الفيروس في ابريل بسبّب الحرَارة	In scientific reading ... the virus is expected to erode in April due to heat.	Academic
متحدث الصحة الحَالَات المُؤكدة حتَى الَان المصَابة بفيروس كورونَا و معضمهَا لبَالغين	Health spokesman has confirmed cases so far infected with coronavirus, mostly for adults.	Media
يَا وزَارة رشوَا البعوض ، هو النَاقل لفيروس كورونَا زَادت الَاصَابَات مع انتشَار البعوض	Ministry of health please spray mosquitoes, as they are carriers of the Coronavirus, increased infections as mosquitoes spread.	Government
خبير صيني يُؤكد ان استنشَاق بخَار المَاء يقتل فيروس كورونَا	A Chinese expert confirms that inhaling water vapor kills coronavirus.	Health professional
علَاج الكرونَا بَاللّيمون و الثوم “رَابط يوتيوب”	Corona treatment with lemon and garlic ”YouTube link”.	Public

Table 6.3: Some examples of false tweets from different source predicted labels

6.7 Conclusion

In this chapter, we identified and analysed one million tweets related to the COVID-19 pandemic in the Arabic language. We performed three experiments which we expect can help to develop methods of analysis suitable for helping Arab World Governments and Public Health Organisations. Our analysis first identifies the topics discussed on social media during the epidemic, detects the tweets that contain false information, and predicts the source of the rumour related tweets based on our previous model for other diseases. The clustered topics are COVID-19 statistics, prayers for God, COVID-19 locations, advise for preventing education, and advertising.

The second contribution in this chapter, is a labeled sample of tweets (2,000 out of 1 million) annotated for false information, correct information, and unrelated. To investigate the replicability and scalability of this annotation, we applied multiple machine learning algorithms with different sets of features. The highest accuracy result was 84% achieved by the LR classifier with count vector set of features and SVC with TF-IDF.

Finally, we also used our previous model to predict the source types of the sampled tweets. Around 60% of the rumour related tweets are classified as written by health professional and academics which shows the urgent need to respond to such fake news. The dataset, including tweet IDs, manually assigned labels for the sampled tweets, and other resources used in this chapter are made freely available for academic research purposes [9].

There are clearly many potential future directions related to analysing social media data on the topics of pandemics. One of the important ways is monitoring the spread of the diseases by finding the infected individuals and defining the infected locations which explained in details in the next Chapter, Chapter 7.

[9] https://doi.org/10.17635/lancaster/researchdata/375

Chapter 7

Monitoring Infectious Diseases

7.1 Introduction

This chapter is based on an extended version of the "Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study" paper (Alsudias and Rayson, 2021b). It explores how adapting for informal language in Arabic Twitter improves monitoring of the COVID-19 pandemic and Influenza epidemic. The aim of this chapter is to analyse Arabic tweets to estimate their usefulness for health surveillance, understand the impact of the informal terms in the analysis, show the effect of the deep learning methods in the classification process, and identify the locations where the infection is spreading. This chapter identifies and creates two Arabic social media datasets for monitoring tweets related to Influenza and COVID-19. It demonstrates the importance of including informal terms, which are regularly used by social media users, in the analysis. It also proves that BERT achieves good results when used with new terms in COVID-19 tweets. Finally, we show that the tweet content may contain useful information to determine the location of the disease spread.

Although millions of items of data appear every day on social media, artificial intelligence through NLP and machine learning algorithms offers the chance to automate their analysis across many different areas, including health. In the area of health informatics and text mining, social media data such as Twitter can be analysed to calculate large-scale estimates of the number of infections, the spread of diseases, or help to predict epidemic events (Joshi, Karimi, Sparks, Paris, and C. Raina Macintyre, 2019b); this field is known as infodemiology, and the systematic monitoring of social media posts and Internet information for public health purposes is known as infoveillance. However, previous research has focussed almost exclusively on English data as we have also shown in Section 5.2 and Section 6.2.

Time is clearly an important factor in the health surveillance domain. In other words, discovering infectious diseases as quickly as possible is beneficial for many organisations and populations as we have seen internationally with COVID-19. It is also important to have multiple independent sources to corroborate evidence of the spread of infectious diseases.

Twitter is one of the main real time platforms that can be used in health monitoring. However, it contains noisy and unrelated information, hence there is a crucial need for information gathering, pre-processing and filtering techniques to discard irrelevant information while retaining useful information. One key task is to differentiate between tweets written for different reasons where someone is infected, or worried about a disease, taking into account the figurative usage of some words related to a disease or spread of infection (Lamb, M. Paul, and Dredze, 2013).

While such tasks are obviously relevant globally, there is little previous research for Arabic speaking countries. There are some characteristics of the Arabic language that make it more difficult to analyse compared with other languages, and NLP resources and methods are less well developed for Arabic than for English. Arabic, which has more than 26 dialects, is spoken by more than 400 million people around the world (Versteegh, 2014). We hypothesise, following our studies in Chapter 4 and Chapter 6, that Arabic speakers will use their own dialects in informal discourse when they express their pain, concerns, and feelings rather than using modern standard Arabic (Hadziabdic and Hjelm, 2014). Table 7.1 describes some examples of Arabic words related to health that may represent different meanings due to dialect differences. For instance, the word (برد) can be understood as influenza in Najdi dialect, and feeling cold in Hejazi dialect (Versteegh, 2014).

The real world motivation of this study in this chapter is to reduce the lag time and increase accuracy in detecting mentions of infectious diseases in order to support professional organisations in decreasing the spread, planning for medicine roll out, and increasing awareness in the general population. We also wish to show that Arabic tweets on Twitter can provide valuable data that may be used in the area of health monitoring by using informal, non-standard and dialectal language, which represents social media usage more accurately.

We focused on COVID-19 and Influenza in particular owing to their rapid spread during seasonal epidemics or pandemics in the Arabic speaking world and beyond. Most people recover within a week or two. However, young children, elderly people, and those with other serious underlying health conditions may experience severe

Word in Arabic	**Potential meaning confusion sets**		
برد	Influenza	feeling cold	
تحصين	Vaccination	reading supplications	
سيلَان	Mucus	nosebleed	
دهَان	Ointment	paint	
رشح	Sneezing	filter the liquid thing	be nominated for a position
مضَاد	Antibiotic	opposite	
حبوب	Tablets	pimples	some kind of food
أَشعة	X-ray	sunlight	
ضعف	Feebleness	double	
مسكن	Painkiller	home	
وصفة	Prescription	method	
فوَار	Medicine	sparkling spring	

Table 7.1: Some examples of Arabic words that have different meaning

complications including infection, pneumonia and death[1]. While it takes specialised medical knowledge to distinguish between the people infected by COVID-19 and Influenza as the symptoms are similar, tracing and planning vaccination and isolation are important for both diseases. In addition, there may be some infected people who do not take the test because of personal concerns and lack of availability of tests in their city, or those who need support to self-isolate.

The overall question being answered in this chapter is how NLP can improve the analysis of the spread of infectious diseases via social media. Our first main contribution is the creation of a new Arabic Twitter dataset related to COVID-19 and Influenza which is labelled with 12 classes, including 11 originating from the Arabic Infectious Disease Ontology, see Chapter 5, and now extended with a new infected category. We used this Ontology since there are no existing medical ontologies such as International Classification of Diseases (ICD) and/or Systematized Nomenclature of Medicine-Clinical Terms (SNOMED) available that originate in Arabic (Joshi, Karimi, Sparks, Paris, and C. Raina Macintyre, 2019b). Crucially, we also showed for the first time the usefulness of informal, non-standard disease related terms using a multi-label

[1] `https://www.who.int`

classification methodology to find personal tweets related to COVID-19 or Influenza in Arabic. We comparatively evaluated our results with and without the informal terms and showed the impact of including such terms in our study. Moreover, we showed the power of machine learning and deep learning algorithms in the classification process. Finally, we developed methods to identify the locations of the infectious disease spread using tweet contents, and this also helped to inform dialect variants and choices.

7.2 Related Work

Previous studies have proven that NLP techniques can be used to analyse tweets for monitoring public health (M. Paul and Dredze, 2011; Aramaki, Maskawa, and Morita, 2011; Breland et al., 2017; Sinnenberg et al., 2017; Charles-Smith et al., 2015; M. Paul, Sarker, et al., 2016). These studies have analysed social media articles that support the surveillance of disease in different languages such as Japanese, Chinese, and English. Diseases that were analysed included listeria, influenza, swine flu, measles, meningitis, and others. We will focus here on previous work related to monitoring influenza and COVID-19 diseases using Twitter data and identifying the location elements from the information on Twitter using NER algorithms.

7.2.1 Influenza Related Research

The Ailment Topic Aspect Model (ATAM) is a model designed by M. J. Paul and Dredze (2012). It uses Twitter messages to measure influenza rates in the United States. It was later extended to consider over a dozen ailments and apply several tasks such as syndromic surveillance and geographical disease monitoring. Similarly, an influenza corpus was created from Twitter (Aramaki, Maskawa, and Morita, 2011). The tweets needed to meet the following two conditions to include them in the training data with infected people and timing: first, the person tweeting, or a close contact is infected with the flu, and second, the tense should be the present tense or recent past.

The goal of a previous study (Lamb, M. Paul, and Dredze, 2013) was to distinguish between flu tweets from infected individuals and others worried about infection in order to improve influenza surveillance. It applied multiple features in a supervised learning framework to find tweets indicating flu. Likewise a sentiment analysis approach was used (Ji, Chun, and Geller, 2016) to classify tweets that included 12 diseases including influenza. A forecasting word model was designed (Hayate, Wakamiya, and Aramaki, 2016) using several words, such as symptoms, that appear in tweets before epidemics to predict the number of patients infected with Influenza.

A previous study (Dai, Bikdash, and Meyer, 2017) used unsupervised methods based on word embeddings to classify health-related tweets. The method achieved an accuracy of 87.1% for the classification of tweets being related or unrelated to a topic. Another study (Hong and Sinnott, 2018) concluded that there is a high correlation between flu tweets and Google Trends data.

A recent survey study (Joshi, Karimi, Sparks, Paris, and C. Raina Macintyre, 2019b) showed how ontologies may be useful in collecting data owing to the structured information they contain. However, there were serious challenges as medical ontologies may consist of medical terms, while the text itself will contain slang terms. The study suggested the inclusion of informal language from social media in the analysis process in order to improve the quality of epidemic intelligence in the future but this was not implemented.

7.2.2 COVID-19 Related Research

Many researchers in computer science have made extensive efforts to show how they can help during pandemics. In terms of NLP and social media, there are various studies that support different languages with multiple goals. These goals include defining topics discussed in social media, detecting fake news, analysing sentiments of tweets, and predicting number of cases (Chandrasekaran et al., 2020).

There have been multiple Arabic datasets published recently (Qazi, Imran, and Ofli, 2020; Shuja et al., 2020). The authors explained the ways of collecting tweets such as time period, keywords, and software library used in the search process, and summed up the statistics for the collected tweets. However, they only included statistical analysis and clustering to generate summaries with some suggestions of future work. Yet, there are some studies with specific goals, such as analysis of the reaction of citizens during a pandemic (Addawood et al., 2020) and identification of the most frequent unigrams, bigrams, and trigrams of tweets related to COVID-19 (Hamoui, Alashaikh, and Alanazi, 2020) . In addition, considering the study by Alanazi et al. (2020) that identified the symptoms of COVID-19 from Arabic tweets, the authors noted the limitation that they used modern standard Arabic keywords only and it would be important to consider dialectical keywords in order to better catch tweets on COVID-19 symptoms written in Arabic because some Arab users post on social media in their own local dialect.

Critically, none of the above studies utilized the Arabic language for monitoring the spread of diseases. There are some Arabic studies that used Twitter with the goal of determining the correctness of health information (Alnemer et al., 2015), analysing health services (Alayba et al., 2017), and proving that Twitter is used by health professionals (Alsobayel, 2016). Moreover, other studies, which did not involve Arabic, used only formal language terminologies when collecting tweets and we would argue that this is not representative of the language usage in social media posts.

7.2.3 Arabic Named Entity Recognition (ANER) Related Research

Previous research on NER aimed to accomplish the following two key goals: (1) the identification of Named Entities and (2) the classification of these entities, usually into coarse-grained categories, including personal names (PER), organisations (ORG), locations (LOC) and dates and times (DATE). Our interest in this chapter was in estimating one of these categories, which is the location element of the information on Twitter. NER methods use a variety of approaches: rule-based, machine learning based, deep learning based and hybrid approaches. These approaches can be used for Arabic, although specific issues arise such as lack of capitalization, nominal confusability, agglutination, and absence of short vowels (Shaalan and Oudah, 2014; Zirikly and Diab, 2015). In addition, there are more challenges in terms of social media content, which includes Arabic dialects and informal terms. There is a lack of annotated data for NER in dialects. The application of NLP tools, originally designed for modern standard Arabic, on dialects leads to considerably less efficiency, and hence we see the need to develop resources and tools specifically for Arabic dialects (Zirikly and Diab, 2015).

The goal of a previous study (Khanwalkar et al., 2013) was to illustrate a new approach for the geo-location of Arabic and English language tweets based on content by collecting contextual tweets. It proved that only 0.70% of users actually use the function of geospatial tagging of their own tweets; thus, other information should be used instead.

7.3 Data Collection and Filtering

There is a lack of an available and reliable Twitter corpora in Arabic in the health domain which makes it necessary for us to create our own corpus. We obtained the data using the Twitter API for the period between September 2019 and October 2020 and collected around 6 million tweets that contained Influenza or COVID-19 keywords.

The keywords are in the code that we will release on GitHub [2]. We collected the tweets weekly since the Twitter API does not allow us otherwise to retrieve enough historical tweets (before the Twitter rules changed as discussed in Section 3.2). We utilised keywords related to Influenza and COVID-19 from the Arabic Infectious Diseases Ontology, see Chapter 5, which includes non-standard terminology. We used a disease ontology because it has been shown to help in finding all the terms and synonyms related to the disease (Ji, Chun, and Geller, 2016).

A previous survey (Joshi, Karimi, Sparks, Paris, and C. Raina Macintyre, 2019b) suggested the inclusion of informal text used in social media in medical ontologies and search processes when collecting data in order to improve the quality of the epidemic intelligence. Therefore, we hypothesise that informal terms may help to find the relevant tweets related to the diseases. Additionally, in the Arabic scenario, we hypothesise that we need to account for dialectal terms.

We filtered the tweets by excluding duplicates, advertisements and spam. Using Python scripts[3], we also cleaned the tweets by removing symbols, links, non-Arabic words, URLs, mentions, hashtags, numbers, and repeating characters. From the resulting dataset, we took a sample of about 4,000 unique tweets, 2,000 tweets on influenza and 2,000 tweets on COVID-19. Then, we used a suite of approaches for pre-processing the tweets, applying the following processes in sequence: Tokenization, Normalization, and Stopword removal. Table 7.2 shows the number of tweets with each label from the ontology after filtering and pre-processing.

7.4 Manual Coding and Inter-coder Reliability

7.4.1 Manual Coding

In order to create a gold standard corpus, our process started with labelling the tweets by two Arabic native speakers, including the author of the thesis, following the guidelines of the annotation process that are described in Appendix C. We manually annotated each tweet with 1 or 0 to indicate Arabic Infectious Diseases Ontology classes which are: Infectious disease name, i.e. Influenza and COVID-19 in our case, Slang term, Symptom, Cause, Prevention, Infection, Organ, Treatment, Diagnosis, Place of the disease spread, and Infected category. We also label each tweet as 1 if the person who wrote the tweet is infected with influenza or COVID-19 and 0 if not.

[2] https://github.com/alsudias

[3] https://github.com/alsudias

Label	**No. of tweets**	
	Influenza	COVID-19
Name of disease	1544	1795
Slang term of disease	456	327
Symptom	398	789
Cause	178	530
Prevention	666	209
Infection	51	15
Organ	2	202
Treatment	152	97
Diagnosis	25	2
Place of the disease spread	17	415
Infected category	52	12
Infected with	907	915

Table 7.2: The number of tweets in each label (each tweet can have multiple labels)

Table 7.3 describes some examples of Arabic Influenza and COVID-19 tweets with their labels.

7.4.2 Inter-coder Reliability

We used Krippendorff's alpha coefficient statistic, which supports multi-label input, to test the robustness of the classification scheme for both datasets (Artstein and Poesio, 2008). The result showed that Krippendorff's alpha score was 0.84 in the Influenza dataset and 0.91 in the COVID-19 dataset which indicates strong agreement between the two manual coders. The remaining disagreement between the annotators was due to informal terms, and Arabic dialects, found in social media. For instance, (البرد لَاعب فينَا لعب), can be understood as cold is playing with us, which represents that an uninfected person or flu is playing with us, indicating an infected person. Another example is (التعَايش مع كورونَا اهون من الحظر), which in English means get along with Corona is easier than the lockdown. This may be classified as an infected person or an uninfected person because the word (التعَايش) has various meanings.

7.5 Methods

In order to create methods to find individuals who have been self-identified as infected and to determine their geo-location in the Twitter dataset, we applied multiple supervised learning algorithms on the labelled dataset and used Named Entity Recognition on the tweet content.

7.5.1 Multi-Label Classification

The overall architecture of our pipeline for finding infected people is shown in Figure 7.1. Using a supervised paradigm, we first annotated the corpus with labelling information as described above, before moving on to classify the tweets by applying machine and deep learning algorithms. We used this method for both the Influenza and COVID-19 case studies. Each tweet has different labels assigned to it. For instance, the first example in Table 7.3 contains the labels: Influenza name (الَانفلونزَا), Slang term of Influenza (زكمة), and Symptom (حرَارة). It also represents that the person is infected with Influenza. Therefore, we assigned a value of 1 to these labels. On the other hand, the tweet does not include the labels: Cause, Prevention, Infection, Organ, Treatment, Diagnosis, Place of the disease spread, and Infected category. Thus, these were marked with 0.

From Table 7.3, we can see that we have a multi-label classification problem where multiple labels are assigned to each tweet. Basically, the following three methods can be used to solve the problem: Problem Transformation, Adapted Algorithm, and Ensemble approaches. For each method, there are different techniques that can be used. We applied the following algorithms, which represent machine learning and deep learning algorithms, to classify the tweets:

Binary Relevance: It treats each label as a separate single class classification problem.

Classifier Chains: It treats each label as a part of a conditioned chain of single-class classification problems, and it is useful to handle the class label relationships.

Label Powerset: It transforms the problem into a multi-class problem with one multi-class classifier that is trained on all unique label combinations found in the training data.

Adapted Algorithm (MLKNN): It is a multi-label adapted K-Nearest Neighbors (KNN) classifier with Bayesian prior corrections.

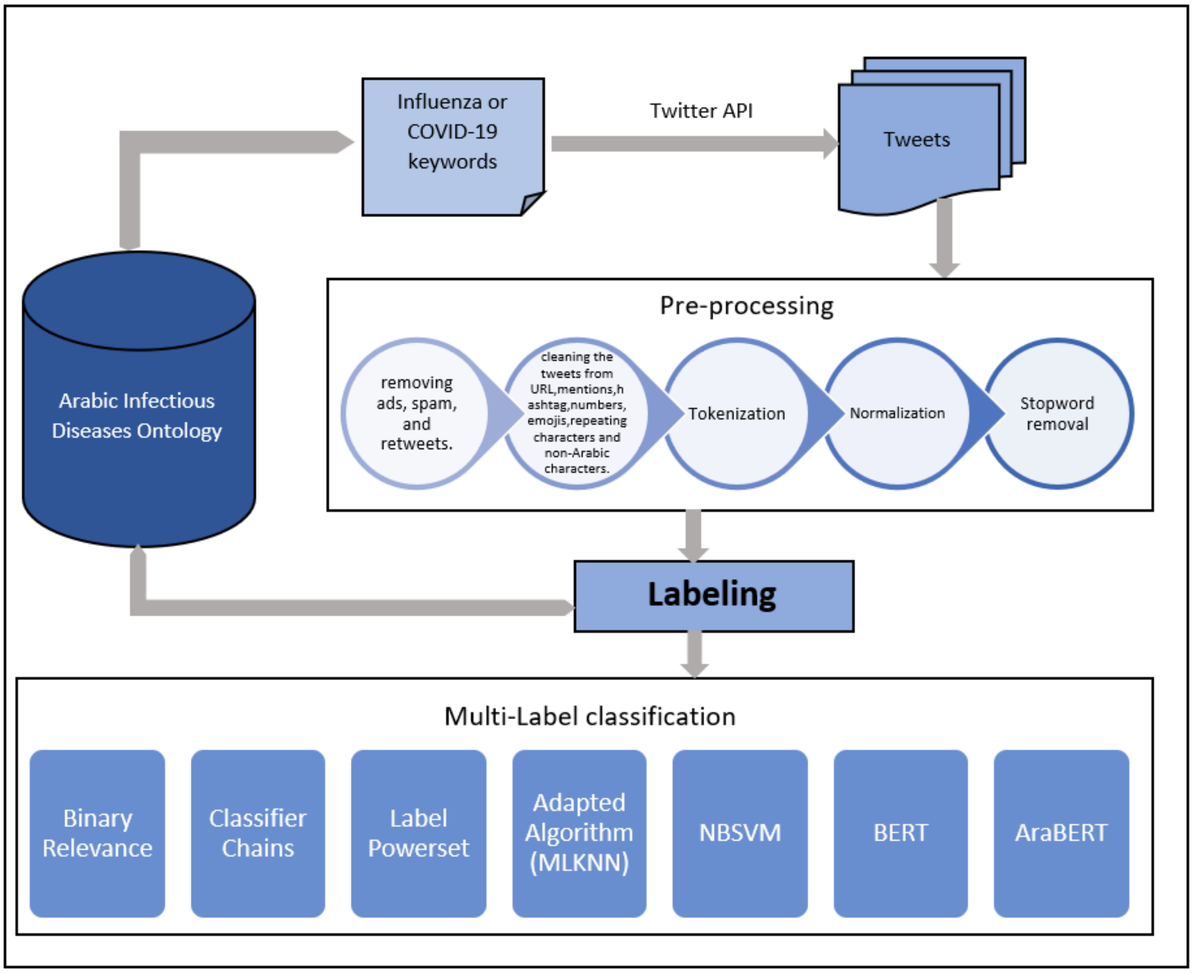

Influenza or COVID-19 keywords
Twitter API
Tweets
Arabic Infectious Diseases Ontology
Pre-processing
removing ads, spam, and retweets.
cleaning the tweets from URL,mentions,hashtag,numbers, emojis,repeating characters and non-Arabic characters.
Tokenization
Normalization
Stopword removal
Labeling
Multi-Label classification
Binary Relevance
Classifier Chains
Label Powerset
Adapted Algorithm (MLKNN)
NBSVM
BERT
AraBERT

Figure 7.1: System architecture.

Support Vector Machine with Naïve Bayes features (NBSVM): It combines generative and discriminant models together by adding NB log-count ratio features to SVM (S. I. Wang and C. D. Manning, 2012).

Bidirectional Encoder Representations from Transformers (BERT): It is a condition where all left and right meaning in both layers to pre-train deep bidirectional representations from unlabelled text (Devlin et al., 2018).

Transformer-based Model for Arabic Language Understanding (AraBERT): It is a pre-trained BERT model designed specifically for the Arabic language (Antoun, Baly, and Hajj, 2020).

Since some labels were 0 for most tweets, we removed these labels in order to avoid overfitting. In other words, we removed the labels that did not appear in most tweets as shown in Table 7.3. The remaining important labels were determined depending on the disease case study because they represented different values for different tweets as justified in Table 7.2. For Influenza, they are Influenza name, Slang term of Influenza, Symptom, Prevention, Treatment, and Infected with. While for COVID-19, they are Name, Slang term of COVID-19, Symptom, Cause, Place, and Infected with. We also repeated the experiment twice to show the effectiveness of the informal terms in the results. One of them had the labels ’disease name’, ’Slang term of Infectious disease’, and ’Infected with’, and the other had all labels, except ’Slang term of Infectious disease’ in both case studies.

In our study, we used the Python scikit-multilearn (Szymański and Kajdanowicz, 2017) and ktrain (Maiya, 2020) libraries and applied different models. To extract the features from the processed training data we used a word frequency approach. We split the entire sample into 75% training and 25% testing sets.

7.5.2 Named Entity Recognition

We followed NER systems that used machine learning algorithms to learn NE tag decisions from annotated text. We used the Conditional Random Fields (CRF) algorithm because it achieved better results than other supervised NER machine learning techniques in previous studies (Zirikly and Diab, 2015).

There were three phases in our geo-location detection algorithm as shown in Figure 7.2. In phase one, the infected person was specified from the multi-label classification algorithm described in the previous section. Then, we retrieved the historical tweets of this person, around 3,000 tweets per person on average, and passed them to the next phase.

The second phase consists of two consecutive stages. First, the tweets were submitted to a Named Entity detection algorithm to select location records from multiple corpora and gazetteers, including ANERCorp (Benajiba and Rosso, 2008; Obeid et al., 2020), and ANERGazet (Benajiba, Rosso, and BenedíRuiz, 2007). A set of location names needs to be filtered out from the general names and ambiguous ones. For example, the word (بَالِي), Bali in English, can be a province in Indonesia or “my mind” as an informal term in Arabic. This step is important in order to ensure that all unrelated location names are not included in the final phase. Second, the identified locations were determined by applying our new Entity Detection Gazetteer which represents Saudi Arabia regions, cities, and district. The data (which will be

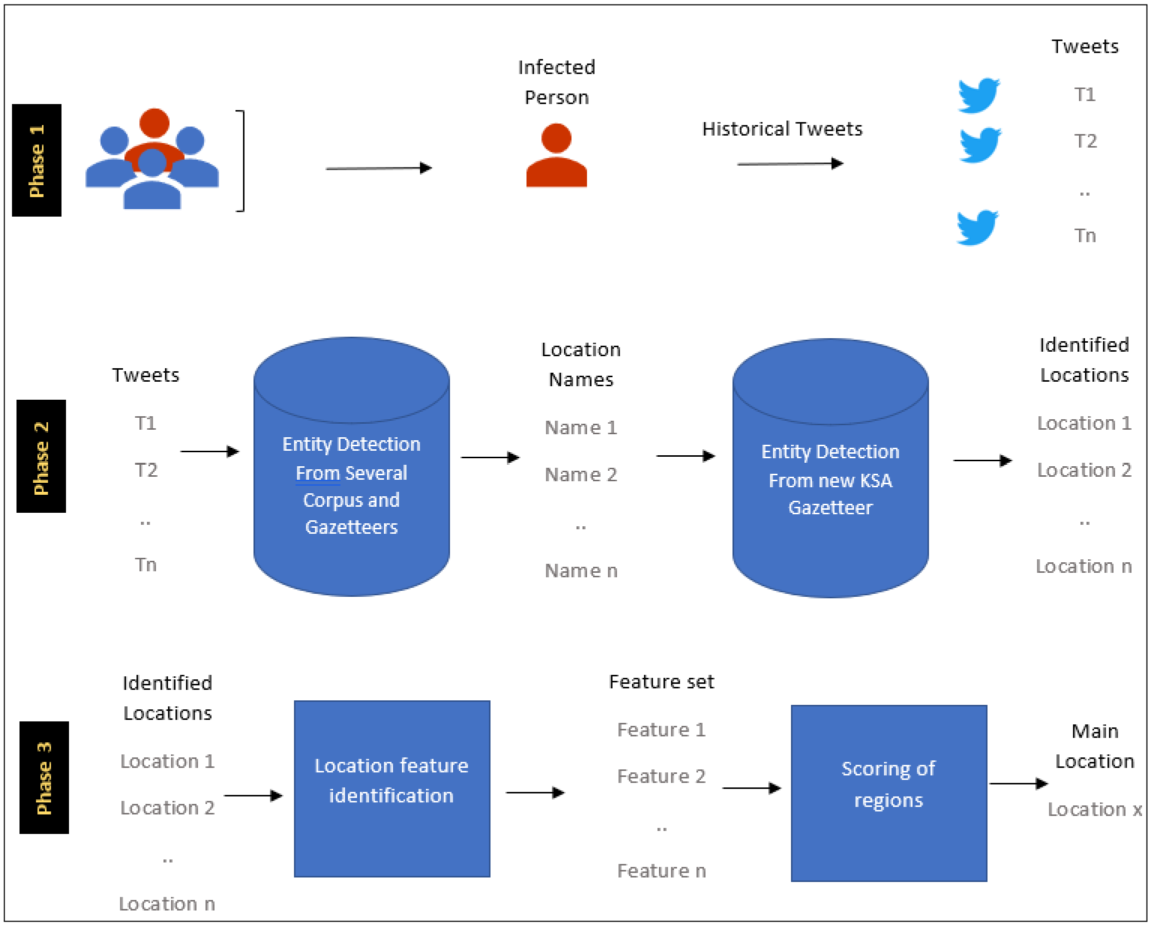


Figure 7.2: Three phases of the Geo-location detection algorithm.

released on GitHub[4]) are public data collected from the Saudi Post website (*National Address Maps* 2020).

In phase three, common features were identified such as the most frequent locations, as well as other features such as occurrence time, which gives a higher score for locations within the last six months. Then, each location is scored by a number, which allows us to rank the list and determine the best estimated main location of the user.

After each tweet set with a predictable location, we compared this location with the location field mentioned in the user account, which is not always set by the

[4] https://github.com/alsudias

user because it is an optional field. Here, we kept only users with valuable location information in either the location or description fields.

7.6 Results

7.6.1 Multi-Label Classification

A multi-label classification problem is more complex than binary and multi-class classification problems. Therefore, various performance measures were calculated to evaluate the classification process such as accuracy, F1-score, recall, precision, Area Under the Receiver Operating Characteristic Curve (ROC AUC), and Hamming loss (X.-Z. Wu and Zhou, 2017). For all these measures, except Hamming loss higher scores are better. For Hamming Loss, smaller values reflect better performance. It is important to note that the accuracy score function in multi-label classification computes only subset accuracy, which means a sample of labels will be taken in the calculation process, as mentioned previously (Szymański and Kajdanowicz, 2017).

Table 7.4 illustrates the performance measures of the seven models on our training dataset with 6, 5, and 3 labels for the Influenza case study. In the 6 labels, which are 'Influenza name', 'Slang term of Influenza', 'Symptom', 'Prevention', 'Treatment', and 'Infected with', the Classifier Chains algorithm achieved the highest results in most measures compared with the other algorithms. It had an F1-score of 86.1%, recall of 81%, precision of 91.8%, AUC of 88.6%, accuracy of 56.2%, and Hamming loss of 8.9%. The Label Powerset algorithm provided a result slightly lower than the Classifier Chains by around 2%. The lowest F1-score was observed for NBSVM which was 58.9%.

The repeated experiment results for the seven models on our training dataset with three labels, which were 'Influenza name', 'Slang term of Influenza', and 'Infected with', are shown in the third part of Table 7.4 and with five labels, which are all labels without 'Slang term of Influenza', are described in the second part of the same table. There was up to 20% enhancement for accuracy in the seven algorithms. The highest F1 score was achieved by the Classifier Chains algorithm which was 88.8%. The recall and precision ranged from 60% to 92%. Consequently, informal terms were shown to represent key factors in the classification process.

Table 7.5 shows the performance measures of the seven models on our training dataset with 6, 5, and 3 labels for the COVID-19 case study. Here, the six labels were different from those in the previous case study because they were determined according to the results from the number of tweets in each label as explained in Table

7.2. The six labels were ‘Name of COVID-19’, ‘Slang term’, ‘Symptom’, ‘Cause’, ‘Place of the disease spread’, and ‘Infected with category’. The best results were achieved by the BERT algorithm with 88.2% F1-score, 86.7% recall, 89.7% precision, 90.3% ROC AUC, 62% Accuracy, and 8.8% Hamming Loss.

The repeated experiment results for the seven models on our training dataset with three labels, which were ‘COVID-19 name’, ‘Slang term of COVID-19’, and ‘Infected with’, are shown in the third part of Table 7.5 and with five labels, which are all labels without ‘Slang term of COVID-19’, are described in the second part of the same table. There was up to 20% enhancement for accuracy in the seven algorithms. The highest F1 score was achieved by the BERT algorithm which was 94.81% followed by AraBERT which was 93.3%. The informal terms in the COVID-19 case study showed around 15% enhancement in the evaluation results.

7.6.2 Named Entity Recognition

A key point to be noted is that our geo-location detection evaluation is based on the location of users where they were tweeting. We filtered tweets that did not have any information in the location field and/or had non plausible locations such as moon or space. We created a manually annotated set from the information in the location field in order to demonstrate greater accuracy. This is due to the ambiguous information in the location field that can be detected by hand. For instance, we found some adjectives of the location like (عروس البحر الاَحمر) and (بوَابة الحرمين) referring to Jeddah city in Saudi Arabia.

In the Influenza study, around 907 users were classified as infected with Influenza, and 397 of these users provided valuable information in their accounts that could be used to identify the location. As a result, our algorithm achieved an accuracy of 45.8% for predicting locations.

Regarding the COVID-19 study, 915 people were considered to be infected and around 358 user accounts had useful information about the location. Therefore, after applying the algorithm, the accuracy was up to 63.6% for identifying the location of the infected users.

7.7 Discussion

To understand the effect of deep learning algorithms on the classification process, we needed to compare the results of the machine learning algorithms with deep learning ones in the two case studies for Influenza and COVID-19. In the Influenza study, the results of deep learning algorithms and the machine learning ones were close to each other. In other words, there was no improvement in the results when applying the deep learning methods such as BERT and AraBERT. On the other hand, in the COVID-19 case study there was up to a 25% enhancement in the results when applying BERT and/or AraBERT. These results helped to confirm that deep learning methods show good returns when dealing with new terms or unknown vocabularies that represent COVID-19 terms.

By applying our previous work, explained in Chapter 4, which classified the sources of the tweets into the following five types: academic, media, government, health professional, and public, we found that informal language was used in the public type (examples 1, 3, and 6 in Table 7.3) while the other types, academic, media, government, and health professional, utilized more formal styles (examples 2, 4, and 5 in the same table). Hence, disease related slang names or other symptoms play an important role in detecting the disease mentions in social media. People not only used slang terms but also expressed their feelings using other terms such as metaphors (Semino et al., 2017). For example, (اهلَا فلونزَا) , which means ‘Hi flu’ shows that the person, who wrote the tweet, was affected by flu. Here 71.9% of the tweets proved that there was a relationship among the informal language used by flu-infected people.

We also found that there was a relationship between ‘Symptom’, ‘Prevention’, and ‘Infected with’ labels. Overall, 64.3% of people infected by influenza sent tweets mentioning symptoms such as sneezing, headache, coughing, or fever. While the tweets about prevention show that 69.3% were written by a person who was not infected with influenza. However, there were a number of tweets that broke these patterns. In other words, we observed tweets written about symptoms that did not represent an infected person or tweets written about prevention that represented an infected person. Table 7.6 shows some examples of the tweets that describe these relationships.

The study by Sarker et al. (2020), which was published recently, proved that users who tested positive for COVID-19 also reported their symptoms using Twitter. Alanazi et al. (2020) described the most common COVID-19 symptoms from Arabic tweets in their study. These symptoms can be further evaluated in clinical settings and used in a COVID-19 risk estimate in near real time.

There are many ways to know the location of the Twitter user such as geo-coordinates, place field, user location, and tweet content. The most accurate method is using the network geo-location system for either the tweet or the user. However, because it is an optional field, less than 3% of users provide this information (Dredze et al., 2013; Qazi, Imran, and Ofli, 2020). In addition, there is noisy information in the user location field because users can type anything like ‘home’ or ‘in the heart of my dad’. As a result, we used the tweet content by assuming that users mentioned helpful information when they tweeted.

On the other hand, some researchers have tried to predict the location of the user using dialect identification from the tweet content (Abdul-Mageed et al., 2020). Although this may prove fruitful, in our scenario it may not reflect the current location that would be required, since a person may tweet in the Egyptian dialect but live in Saudi Arabia.

7.8 Conclusion

This chapter has, for the first time, shown that Arabic social media data contain a variety of suitable information for monitoring of Influenza and COVID-19, and crucially, it has improved on previous research methodologies by including informal language and non-standard terminology from social media which have been shown to help in filtering unrelated tweets. It should be noted that we are not trying to provide a single source of information for public health bodies to use, but want to provide a comparable information source through which to triangulate and corroborate estimates of disease spread against other more traditional sources.

We also introduced a new Arabic social media dataset for analysing tweets related to Influenza and COVID-19. We labelled the tweets for categories in the Arabic Infectious Disease Ontology which includes non-standard terminology. Then, we used multi-label classification techniques to replicate the manual classification. The results showed a high F1 score for the classification task and showed how non-standard terminology and informal language are important in the classification process with an average improvement of 8.8%. The dataset, including tweet IDs, manually assigned labels, and other resources used in this chapter are released freely for academic research purposes, with a DOI via Lancaster University’s research portal [5].

[5] https://bit.ly/3DVvE8G

Moreover, we applied a Named Entity Recognition algorithm on the tweet content to determine the location and spread of the infection. Although the number of users was limited, the results showed good accuracy in the analysis process.

The thesis comes to a close with a rundown of the research conducted and its conclusions in the next chapter, Chapter 8, which is the last one. This will include going through the research questions that were asked at the start of the study and evaluating how well they have been answered.

Tweet in Arabic	**Tweet in English**	Name	Slang term	Symptom	Cause	Prevention	Infection	Organ	Treatment	Diagnosis	Place	Infected category	**Infected with**
و الحل مع الَانفلوَانزَا فوق الحرَارة زكمه ذبحتنَى ذبح	What is the solution with flu, fever and cold killed me	1	1	1	0	0	0	0	0	0	0	0	**1**
حملة تطعيم ضد الَانفلونزَا بَالتعَاون مع مستشفَى الملك خَالّد	Influenza vaccination campaign in cooperation with King Khalid Hospital.	1	0	0	0	1	0	0	0	0	0	0	**0**
صبَاح الَانفلونزَا	Flu morning	1	0	0	0	0	0	0	0	0	0	0	**1**
تجربتي بعد تَاكد اصَابتي بكوفيد٩١ كَانت الَاعرَاض طفيفة لَاحظت ان الفيروس يعمل علَى مرَاحل في البدَاية لَاحظت التعرق و الصدَاع ثم الم العين	My experience after my infection with Covid-19 was confirmed, the symptoms were slight initially, I noticed that the virus works in stages, at first I noticed sweating, headache, and then eye pain.	1	0	1	1	0	0	1	0	0	0	0	**1**
غسل اليدين بَالصَابون و لبس الكمَامة الطبية الكم عدّ من الَاجرَاءَات الَاحترَازية اللّتي لَا تزَال افضل السبل للّوقَاية من كورونَا	Washing hands with soap, and wearing a medical mask ... Here are a number of precautionary measures that are still the best ways to prevent Corona	0	1	0	0	1	0	0	0	0	0	0	**0**
اول مَا ادهشني هو الخمول و الم في العظَام، وصدَاع غربب ثم الَاسهَال. لم اكن اتوقع كورونَا لَان الَاعرَاض كَانت خفيفة.	The first thing that struck me was lethargy, pain in the bones, a strange headache, and then had diarrhea. I didn't expect Corona cause the symptoms were mild.	1	0	1	0	0	0	1	0	0	0	0	**1**

Table 7.3: Examples of tweet with their assigned labels (1 or 0)

No. of Labels	Multi-label classification technique	F1-Score (%)	Recall (%)	Precision (%)	ROC AUC (%)	Accuracy (%)	Hamming Loss (%)
6	Binary Relevance	73.1	74.4	71.9	79.7	39.6	18.7
	Classifier Chains	86.1	81	91.8	88.6	56.2	8.9
	Label Powerset	85.7	83.8	87.6	88.7	56.2	9.7
	Adapted Algorithm (MLKNN)	76.9	75.5	78.4	82.3	39.9	15.5
	BERT	78.1	83.4	73.4	85.4	38.9	13.7
	AraBert	79.7	72.7	88.2	83.9	49.2	12.5
	NBSVM	58.9	46.3	81.2	70.9	26.8	18.9
5	Binary Relevance	75.5	76.9	74.1	80.7	45.1	18.3
	Classifier Chains	88	85.5	90.5	90.2	64.9	8.5
	Label Powerset	87.6	86.2	89.2	90	63.9	8.9
	Adapted Algorithm (MLKNN)	79.9	76.4	83.9	84	47.9	14
	BERT	84.1	83.1	85	88	57.5	10.3
	AraBert	87.3	86.3	88.4	90	64.3	9
	NBSVM	61.6	49.7	81.2	72	26.8	20.2
3	Binary Relevance	80.8	80	81.7	81.2	60.4	18.8
	Classifier Chains	88.8	85.7	92.2	89.3	72.4	10.7
	Label Powerset	88.3	88	88.6	88.4	70.8	11.6
	Adapted Algorithm (MLKNN)	80.9	84.7	77.5	80.2	54	19.8
	BERT	87.6	93.9	82.1	88.9	68.1	11.7
	AraBert	85.9	81.5	90.9	86.8	66.9	13.1
	NBSVM	79.5	75.1	84.3	82.1	59.9	17.1

Table 7.4: Training results of the seven algorithms with 6, 5, and 3 labels for the Influenza case study

No. of Labels	Multi-label classification technique	F1-Score (%)	Recall (%)	Precision (%)	ROC AUC (%)	Accuracy (%)	Hamming Loss (%)
6	Binary Relevance	54.6	52.8	56.6	64	15.6	33.3
	Classifier Chains	53.9	49.8	59.7	64.2	18.5	32.3
	Label Powerset	58.6	59.4	57.9	66.5	22.2	31.8
	Adapted Algorithm (MLKNN)	54.5	51	58.4	64.4	10	32.4
	BERT	88.2	86.7	89.7	90.3	62	8.8
	AraBert	82	84.4	79.8	86	50.5	13.6
	NBSVM	64.3	51.7	85	73.1	20.7	21.7
5	Binary Relevance	57	56	58.1	63.1	15.8	35.9
	Classifier Chains	56.2	53	59.9	63.3	18.3	35.1
	Label Powerset	60.8	63.4	58.4	65	22	34.8
	Adapted Algorithm (MLKNN)	56.5	54.6	58.7	63.1	10.4	35.7
	BERT	87.3	87.9	86.7	88.9	59	10.9
	AraBert	86.3	92.7	80.7	88.6	53.9	12.1
	NBSVM	55.2	40.6	86.4	67.9	17.9	28
3	Binary Relevance	68.5	69	68	69.2	36.9	30.8
	Classifier Chains	69.7	68.1	71.4	71.2	39.9	28.3
	Label Powerset	70.3	69	71.5	71.6	40.1	28.3
	Adapted Algorithm (MLKNN)	71.6	70.7	72.6	72.8	41.4	27.1
	BERT	94.8	96.4	93.3	94.9	93.2	5.1
	AraBert	93.3	94.8	91.9	93.5	85.3	6.5
	NBSVM	70.6	59.6	86.5	75.4	46.5	24.2

Table 7.5: Training results of the seven algorithms with 6, 5, and 3 labels for the COVID-19 case study

Tweet in Arabic	Tweet in English	Description
يَا شين صدَاع الَانفلوَانزَا	Flu headache is bad	The relationship between symptom and infected with influenza
شكلي بموت من الَانفلوَانزَا من صحيت عطست ١٠ مرَات	I think I will die from flu, I sneeze 10 times from the time I wake up	The relationship between symptom and infected with influenza
لقَاح الَانفلوَانزَا لَا يمنع التهَابَات الرشح كمَا يعتقد البعض ولكنه يمنع التهَابَات انفلوَانزَا أ و ب الرِئوية الخطيرة التي تقتل أعدَاد كبيرة حول العَالم	The flu vaccine does not prevent colds, as some believe, but it prevents serious influenza A and B infections that kill large numbers around the world.	The relationship between prevention and non-infected with influenza
وش سويتي فيني يَا كورونَا اسبوعين للّحين ومو قَادرة احس بطعم شي	Corona, what did you do for me? For two weeks, I will not be able to feel the taste of something	The relationship between symptom and infected with COVID-19
الريَاض تسجل ٣٢٠ اصَابة جديدّة بكورونَا و ١٥ حَالة وفَاة	Riyadh records 320 new coronavirus cases and 15 deaths	The relationship between place and non-infected with COVID-19
التزموَا بَالَاحترَازَات و الوقَاية من كورونَا فَالموجة فعلَا بدَات، فَارتدوَا الكمَامَات وَابتعدوَا عن التجمعَات وتعقموَا وَاغسلوَا ايديكم بَالمَاء وَالصَابون لمدة ثلَاثين ثَانية	Adhere to the precautions and prevention from Corona, as the wave has really started, so wear masks, stay away from gatherings, sterilize and wash your hands with soap and water for a period of no less than thirty seconds	The relationship between prevention and non-infected with COVID-19

Table 7.6: Some Examples of tweets describing relationships between symptom, prevention, and infected with labels

Chapter 8

Conclusion

8.1 Summary of Thesis

In this thesis, we have analysed the spread of infectious diseases in Arabic speaking countries, with a focus on Saudi Arabia via Twitter. This required advancements in various fields. First, it required the creation of Arabic Twitter dataset. Second, we needed to create a classification system that can find the source of information. Third, it was necessary to build an Arabic Infectious Diseases Ontology. Fourth, we needed to create methods to analyse tweets during a pandemic in both qualitative and quantitative ways including identifying the topics discussed on social media during the epidemic, detecting the tweets that contain false information, and predicting the source of the rumour related tweets. Fifth, it was required to create a system that measures the spread of infectious diseases. Finally, we also needed to create a system that can predict the approximate location of the infected person. We addressed all points in this thesis which is split into eight chapters as shown in the following.

Chapter 1 covered the aims of this research, which are using Arabic Twitter content to measure the incidence of infectious disease, and the contributions of this thesis. It also offered a concise introduction to the research subject, the motivation of the work, and brief overview of Arabic language, infectious diseases, and Twitter.

Chapter 2 provided background information about infectious diseases, Arabic natural language processing, analysing social media data, and machine learning algorithms which are the key work topics of this thesis. It also included the current and previous work in the field of analysing health-related Arabic and non-Arabic Twitter data, as well as determining geo-location from tweets.

In Chapter 3, there was an explanation of the data collection process for 21 infectious diseases shown in Table 3.1. It focused on Twitter rules, ethical approval, methodology, some statistics, and a discussion of potential future possibilities within the dataset. This chapter achieved objective 1, which was the creation of a new Arabic infectious diseases dataset.

Chapter 4 showed the first experiment in classifying information sources in Arabic Twitter to support online monitoring of infectious diseases. The goal of this chapter was to investigate how taking into account the source of information as well as the content of tweets might help with dependability and trust judgements. It classified tweets into five types of sources: academic, media, government, health professional, and public. We performed two experiments. First, native speakers judge whether they could manually classify tweets into these five groups, and then we repeated the experiment using various machine learning classifiers. We found that inter-annotator agreement was 0.84 for this task, and machine learning classifiers were able to correctly identify the type of source of a tweet with 77.2% accuracy without knowledge of the user and their bio or profile, but with 99.9% accuracy when provided with this information. Objective 2, which was to create and evaluate a classification system to find the source of information, was carried out in this chapter.

In Chapter 5, the Arabic Infectious Disease Ontology was developed. It included the scientific vocabularies of infectious diseases with their informal equivalents. It was written in the Arabic language and was the first ontology in Arabic specialising in the infectious disease domain. It contained 11 classes, 21 object properties, 11 datatype properties, and 215 individual concepts. The goal of this ontology is to support the analysis of the spread of infectious disease in any Arab Country. It has a wide range of applications in academia and the real world, including Question Answering on the Semantic Web and, in our case, online monitoring of health events. This chapter completed Objective 3, which was creating and evaluating an Arabic Infectious Diseases Ontology.

Chapter 6 presented an analysis study of tweets during the pandemic, COVID-19. It analysed them in three different ways: identifying the topics discussed during the period, detecting rumours, and predicting the source of the tweets. The clustered topics were COVID-19 statistics, prayers for God, COVID-19 locations, advice for preventing education, and advertising. We sampled 2,000 tweets and labelled them manually as false information, correct information, and unrelated. Then, we applied three different machine learning algorithms, Logistic Regression, Support Vector Classification, and Naïve Bayes with two sets of features, word frequency approach, and word embeddings. We found that machine learning classifiers are able to correctly identify the rumour related tweets with 84% accuracy. We also tried to predict the source of the rumour related tweets depending on our previous model done in Chapter 4, which was about classifying tweets into five categories: academic, media, government, health professional, and public. Around 60% of the rumour related tweets were classified as written by health professionals and academics. The results of this study will be useful for Public Health Organisations and Arab World Governments. Objective 4, applying different quantitative and qualitative studies to analyse tweets during the COVID-19 pandemic, was achieved in this chapter.

Chapter 7 introduced a new approach for monitoring the COVID-19 pandemic and Influenza epidemic with adaptation for informal language in Arabic Twitter data. It also expanded the scope of the research to predict the location of the infected person via a tweet's content. The goal of this chapter was to analyse Arabic tweets to estimate their usefulness for health surveillance, understand the impact of the informal terms in the analysis, show the effect of deep learning methods in the classification process, and identify the locations where the infection is spreading. We applied the following multilabel classification techniques: Binary Relevance, Classifier Chains, Label Powerset, Adapted Algorithm (multilabel adapted k-Nearest Neighbors [MLKNN]), Support Vector Machine with Naïve Bayes features (NBSVM), Bidirectional Encoder Representations from Transformers (BERT), and AraBERT (Transformer-based Model for Arabic Language Understanding) to identify tweets appearing to be from infected individuals. We also used Named Entity Recognition to predict the place names mentioned in the tweets. We achieved an F1 score of up to 88% in the Influenza case study and 94% in the COVID-19 one. Adapting for nonstandard terminology and informal language helped to improve accuracy by as much as 15%, with an average improvement of 8%. Deep learning methods achieved an F1 score of up to 94% during the classifying process. Our geolocation detection algorithm had an average accuracy of 54% for predicting the location of users according to tweet content. This chapter demonstrated the importance of including informal terms, which are regularly used by social media users, in the analysis. It also proved that BERT achieves good results when used with new terms in COVID-19 tweets. Finally, the tweet content may contain useful information to determine the location of disease spread. This chapter accomplished Objectives 5 and 6, which were to create a system that measures the spread of infectious disease and also to be able to identify the approximate location of the infected people.

8.2 Research Questions Revisited

Now we will look at how the work done so far has answered the research questions posed in Section 1.3, which we'll repeat here for convenience.

1. **How can we differentiate the sources of the information on social media in the area of health surveillance?**

This research question has been addressed in Chapter 4. We highlighted the important issues related to levels of trust, quality, and reliability of the information online by considering the variety of information sources especially for health related topics. Here, we were able to classify tweets into the five categories: academic, media, government, health professional, and public, with favourable accuracy on the tweet content itself, and high accuracy using the bio information from the user. However, in emergency situations and fast moving scenarios, it is important to understand the veracity of information released in social media, in order to avoid acting on false information.

2. **In what way does the Infectious Diseases Ontology affect the collection and analysis of the spread of infectious disease?**

   In Chapter 5, this research question has been answered. We showed that there was plenty of research that proved that the ontologies gave good results in health surveillance research in section 5.2. However, until we started our research and to our knowledge, there was not an existing ontology in the Arabic language and infectious disease domain. As a result, there was a need to build our own one. After reviewing the existing ontologies and studying the goal of the overall research question, we determined the ontology architecture based on a hybrid approach to achieve higher quality as shown in Figure 5.1. It included various ordered stages: 1) Data Sources Selecting and Gathering, 2) Pre-processing, 3) Term Extraction, 4) Finding Synonyms, 5) Adding Informal Terms, 6) Obtaining Concepts, and 7) Defining Relations and Updating them. We explained in detail all these phases in the chapter and discussed the reasons for doing them. To determine the slang terms related to infectious diseases, a questionnaire was sent to different native speakers. We undertook various evaluations of the ontology by multiple techniques including a domain expert evaluation to check medical correctness. This ontology would be used in the next studies related to RQ3, RQ4, and RQ5. It would be useful to connect our Ontology to an Upper Arabic ontology in order to expand its knowledge and usage. Some other potential directions for future work include expanding the Ontology to other diseases in order to build an Arabic Ontology that covers all kinds of diseases.

3. **What are the topics that are discussed during pandemics?**

   This research question was addressed in Chapter 6. While much research was performed related to the recent pandemic, COVID-19, few of them were focused on the Arabic language as shown in Section 6.2. This chapter was derived from the paper presented and published at the NLP COVID-19 Workshop, an emergency workshop at ACL 2020[1]. In this chapter, we performed three experiments which we expect could help to develop methods of analysis suitable for assisting Arab World Governments and Public Health Organisations. Our analysis identified the topics discussed on social media during the epidemic, detected the tweets that contain false information, and predicted the source of the rumour related tweets. Since false information has the potential to play a dangerous role in topics related to health, there is a need to enhance and automate the automatic detection process supporting different languages beyond just English.

4. **How can we measure the spread of infectious diseases in Arabic speaking countries, with a focus on Saudi Arabia via Twitter?**

   In Chapter 7, this research question has been addressed. We performed the analysis in two cases studies, COVID-19 and Influenza. We first introduced an Arabic social media data set for analysing tweets related to Influenza and COVID-19. We labeled the tweets for categories in the Arabic Infectious Disease Ontology, which includes nonstandard terminology, that was built in response to RQ3. Then, we used multilabel classification techniques to replicate the manual classification. The results showed a high F1 score for the classification task and showed how nonstandard terminology and informal language are important in the classification process, with an average improvement of 8.8%. It demonstrated the importance of including informal terms, which are regularly used by social media users, in the analysis. It also proved that BERT achieves good results when used with new terms in COVID-19 tweets. Within these results, it could be useful to enhance the performance of the system, including expanding the data used to train the classifier and analysing different infectious diseases.

5. **How is it possible to identify the approximate location of the infected people using content of the tweet?**

   This research question was answered also in Chapter 7. We used Named Entity Recognition to predict the place names mentioned in the tweets after identifying the infected people. There were three phases in our geolocation detection algorithm as shown in Figure 7.2. In phase 1, the infected person was specified from the multilabel classification algorithm described in the previous study,

[1] https://www.nlpcovid19workshop.org/

answering RQ4. Then, we retrieved the historical tweets of this person and passed them to the next phase. In phase 3, common features were identified, such as the most frequent locations, as well as other features, such as occurrence time, which gives a higher score for locations within the last six months. Then, each location is scored by a number, which allows us to rank the list and determine the best estimated main location of the user. This chapter proved that the tweet content may contain useful information to determine the location of disease spread. Although the number of users was limited, the results showed good accuracy in the analysis process.

8.3 Achieved Contributions

In this research, the contributions that have been achieved are:

- The creation of new Arabic infectious diseases dataset[2].
- The creation of a new Arabic Infectious Diseases Ontology[3].
- The creation and evaluation of a classification system to find the source of information[4].
- A study of quantitative patterns and qualitative themes of how people talk about infectious diseases in Arabic social media[5].
- An analytical study of the spread of infectious diseases by adapting for informal language[6].
- Using Arabic Named Entity Recognition (ANER) methods to predict the location of the user depending on the content of the tweet[7].

[2] https://github.com/alsudias/Arabic-Infectious-Diseases-Twitter-Dataset

[3] https://www.research.lancs.ac.uk/portal/en/datasets/arabic-infectious-disease-ontology(39dbef60-ae9b-4405-99c8-35a41e95a3e0).html

[4] https://www.research.lancs.ac.uk/portal/en/datasets/arabic-tweets-about-infectious-diseases(68b307a8-510b-42f6-93af-f0cc7d9a75a1).html

[5] https://www.research.lancs.ac.uk/portal/en/datasets/covid19-arabic-tweets(400f5a2a-a34e-4d3b-a72d-48c2d0df0513).html

[6] https://www.research.lancs.ac.uk/portal/en/datasets/influenza-and-covid19-arabic-labeled-tweets(ff9222a0-eb68-4fa4-9171-4a186de6c14c).html

[7] https://github.com/alsudias

8.4 Future Work

This thesis prompts many further directions which can be followed on Arabic social media analysis. The advancement of Arabic NLP research and the availability of more Arabic tools and resources are both critical to improving present Arabic analysis techniques and approaches. In other words, researchers on Arabic NLP now have resources, tools, and results that can be used to advance the research on Arabic social media data related to infectious disease. However, they can address the limitations of this study in different directions.

This study paves the way for further research on the Arabic language and social media data or any other language, especially infectious disease related ones. The following are some prospective areas of future research.

- Exploring alternative sources of informal Arabic data, such as WhatsApp and Instagram apps, blogs, and YouTube comments, to cover the majority of the sources, as well as employing voice recognition on spoken Arabic to develop a corpus that includes several sources of the Arabic dataset.
- Comparing different infectious diseases with either on historical datasets such as (Ahmed, Peter Bath, et al., 2018) or future ones to predict (Molaei et al., 2019).
- Improving the result of classification by applying various features and using multiple machine learning algorithms like deep learning models.
- Building an annotation tool that can facilitate the annotation process and reduce the cost similar to the work performed in (Saleh and Al-Khalifa, 2009)
- Extending the study to additional dialectal languages, such as German, Italian, and Chinese[8]. It can be also useful to analyse the effect of slang terms in any language, including English, in many applications.
- Extending our ontology with other diseases and connecting it to an Upper Arabic ontology in order to expand its knowledge and usage.
- Developing a system that can detect the misinformation and disinformation in Arabic social media especially that related to health. Suarez-Lledo and Alvarez-Galvez (2021) and Yuxi Wang et al. (2019) proved the need to develop the research in the area of the spread of misinformation on social media particularly infectious disease and vaccines.

[8] `https://en.wikipedia.org/wiki/Dialect`

- With more Arabic tweets available on the web in different topics, there is a need to build a real-time classifier related to public health information. A survey study by Jordan et al. (2019) analysed several studies related to health surveillance and provided diverse future direction within Twitter data and a study by Zubiaga, Spina, et al. (2015) provided an example of an automatic classifier of trending topics in Twitter.

Appendix A

Collection Data

A.1 Ethics application

Here is the ethical approval application as approved by Lancaster University FST Ethics Committee[1].

[1] https://www.lancaster.ac.uk/sci-tech/research/ethics

Faculty of Science and Technology Research Ethics Committee (FSTREC)
Lancaster University

Application for Ethical Approval for Research

This form should be used for all projects by staff and research students, whether funded or not, which have not been reviewed by any external research ethics committee. If your project is or has been reviewed by another committee (e.g. from another University), please contact the FST research ethics officer for further guidance.

In addition to the completed form, you need to submit **research materials** such as:

i. Participant information sheets
ii. Consent forms
iii. Debriefing sheets
iv. Advertising materials (posters, e-mails)
v. Letters/emails of invitation to participate
vi. Questionnaires, surveys, demographic sheets that are non-standard
vii. Interview schedules, interview question guides, focus group scripts

Please note that **you DO NOT need to submit pre-existing questionnaires or standardized tests** that support your work, but which cannot be amended following ethical review. These should simply be referred to in your application form.

Please submit this form and any relevant materials **by email as** a **SINGLE attachment** to *fst-ethics@lancaster.ac.uk*

Section One

Applicant and Project Information

Name of Researcher:
Lama Alsudias
Project Title:
Using Twitter to support analysis of the spread of Infectious Diseases in Saudi Arabia
Level: Masters, PhD, Staff
PhD
Supervisor (if applicable):
Dr. Paul Rayson
Researcher's Email address: ***l.alsudias@lancaster.ac.uk***
Telephone:
Address: ***InfoLab21, Lancaster University, Lancaster, LA1 4WA, UK.***

Names and appointments/position of all further members of the research team:

Is this research externally funded? If yes,

ACP ID number:
Funding source:
Grant code:

Does your research project involve any of the following?

☐ Human participants (including all types of interviews, questionnaires, focus groups, records relating to humans, use of internet or other secondary data, observation etc.)

☐ Animals - the term animals shall be taken to include any non-human vertebrates or cephalopods.

☐ Risk to members of the research team e.g. lone working, travel to areas where researchers may be at risk, risk of emotional distress

☐ Human cells or tissues other than those established in laboratory cultures

☐ Risk to the environment

☐ Conflict of interest

☐ Research or a funding source that could be considered controversial

☑ Social media and/or data from internet sources that could be considered private

☐ any other ethical considerations

Yes – complete the rest of this form

No – your project does not require ethical review or submission of this form

Section Two

Type of study

☐ Includes *direct* involvement by human subjects. ***Complete all sections apart from Section 3.***

☒ Involves *existing documents/data only*, or the evaluation of an existing project with no direct contact with human participants. ***Complete all sections apart from Section 4.***

If your research involves data from chat rooms and similar online spaces where privacy and anonymity are contentious, please complete all sections

Project Details

1. Anticipated project dates (month and year)
Start date: End date:
October 2018 September 2021

2. Please briefly describe the background to the research (no more than 150 words, in lay-person's language):

In general, social media can be useful as a source of health information. Many government organisations, academic experts, professions, and media outlets in the field of health and medicine release useful health information needed by the public. Due to its popularity as a communication platform for real time messages, it could potentially be useful to analyse tweets for health surveillance purposes. We hypothesise that this may reduce reporting lag time compared to the traditional approaches for monitoring infectious diseases.

3. Please state the aims and objectives of the project (no more than 150 words, in lay-person's language):

Monitoring the spread of infectious diseases is important using many approaches, formal and informal, due to their effect on human life. Formal strategies such as hospital records, laboratory statistics, or household surveys need time to produce results. On the other hand, informal methods like search engine query logs or social media may detect diseases near to real time. Governments now make extensive efforts to detect infectious diseases as soon as possible in order to provide medicines and prevent the spread of the disease. So, there is need for a system that enhances the speedier detection of disease in the future. The analysed tweets will be used to support better understanding of the spread of infectious Diseases in the Arabic language.

4. Methodology and Analysis:

Our extension to the previous work will focus on how to build a new Infectious Diseases Ontology in Arabic with informal and formal terms. A new Arabic dataset will be created via Twitter API and depends on keywords from the new Ontology with the location of tweets to be Saudi Arabia. The machine Learning algorithms with Natural Language Processing (NLP) techniques especially for Arabic will be used for the classification.

Section Three

Secondary Data Analysis

Complete this section if your project involves *existing documents/data only*, or the evaluation of an existing project with no direct contact with human participants

1. Please describe briefly the data or records to be studied, or the evaluation to be undertaken.

Twitter data written in Arabic that mentions any information about infectious diseases. We will collect the tweets via a small set of keywords and hashtags related to infectious diseases. For example, Tweets that are written during an outbreak or discuss diseases in general. We will then be able to compare the quantitative profile of these with other published data on disease outbreaks as mentioned above. The scale of our data will be restricted by the Twitter API restrictions on how much can be collected per API call, or over time.

2. How will any data or records be obtained?
Via the Twitter API which is a freely available software interface from Twitter through which we can computationally script the data collection and allows us to collect tweets within the last 7 days for any public account.

3. Confidentiality and Anonymity: If your study involves re-analysis and potential publication of existing data but which was gathered as part of a previous project involving direct contact with human beings, how will you ensure that your re-analysis of this data maintains confidentiality and anonymity as guaranteed in the original study?
We have contacted Twitter to obtain a developer licence for accessing the API. Their terms of service allow us to publish only the tweet ID, which is the identification number for tweets, and this will not contain the text itself. Publishing a list of tweet IDs for the whole corpus is standard practice in the NLP community to enable replicability. For publication purposes of example tweets, we will not reveal specific Twitter usernames and we will paraphrase the tweet itself so that anonymity can be maintained i.e. a specific tweet and user cannot be recovered from examples in papers via a web search query. We have also recently been in touch directly with a senior policy strategist at Twitter HQ to check that our procedures for paraphrasing tweets meet their T&Cs. This practice has already been approved by the FHM Research Ethics Committee for another student in DHR/FHM co-supervised by Paul Rayson.

4. What plan is in place for the storage of data (electronic, digital, paper, etc)? Please ensure that your plans comply with the General Data Protection Regulation (GDPR) and the (UK) Data Protection Act 2018.
An Excel file that contain the tweets will be shared with me and my supervisor only during the PhD. This will be stored on encrypted password protected laptops and on Linux virtual machines in ISS's cluster during collection.

5. What are the plans for dissemination of findings from the research?
The findings of this research will be part of the PhD thesis. I will also present some result as papers to workshops, journals, or conferences in the NLP community e.g. ACL, LREC, TACL, CL. Findings will only include quantitative summaries of data and paraphrased example tweets as described above.

6a. Is the secondary data you will be using in the public domain? YES
6b. If NO, please indicate the original purpose for which the data was collected, and comment on whether consent was gathered for additional later use of the data.

7.What other ethical considerations (if any), not previously noted on this application, do you think there are in the proposed study? How will these issues be addressed?
No

8a. Will you be gathering data from discussion forums, on-line 'chat-rooms' and similar online spaces where privacy and anonymity are contentious? NO

If yes, your project requires full ethics review. Please complete all sections.

Section Four

Lancaster University

Participant Information

Complete this section if your project includes *direct* involvement by human subjects.

1. Please describe briefly the **intended human participants** (including number, age, gender, and any other relevant characteristics):

2. How will participants be **recruited** and from where?

3. Briefly describe your **data collection methods**, drawing particular attention to any potential ethical issues.

4. Consent
4a. Will you take all necessary steps to **obtain the voluntary and informed consent** of the prospective participant(s) or, in the case of individual(s) not capable of giving informed consent, the permission of a legally authorised representative in accordance with applicable law? **YES/ NO**
If yes, please go to question 4b. If no, please go to question 4c.

4b. Please explain the procedure you will use for **obtaining consent**?. If applicable, please explain the procedures you intend to use to gain permission on behalf of participants who are unable to give informed consent.

4c. If it will be necessary for participants to take part in the study **without their knowledge and consent at the time**, please explain why (for example covert observations may be necessary in some settings; some experiments require use of deception or partial deception – not telling participants everything about the experiment).

5. Could participation cause **discomfort** (physical and psychological eg distressing, sensitive or embarrassing topics), **inconvenience or danger beyond the risks encountered in normal life**? Please indicate plans to address these potential risks. State the timescales within which participants may withdraw from the study, noting your reasons.

6. How will you protect participants' **confidentiality and/or anonymity** in data collection (e.g. interviews), data storage, data analysis, presentation of findings and publications?

7. Do you anticipate any ethical constraints relating to **power imbalances or dependent relationships**, either with participants or with or within the research team? If yes, please explain how you intend to address these?

8. What potential **risks may exist for the researcher** and/or research team? Please indicate plans to address such risks (for example, noting the support available to you/the researcher; counselling

considerations arising from the sensitive or distressing nature of the research/topic; details of the lone worker plan you or any researchers will follow, in particular when working abroad.

9. Whilst there may not be any significant direct **benefits to participants** as a result of this research, please state here any that may result from participation in the study.

10. Please explain the **rationale for any incentives/payments** (including out-of-pocket expenses) made to participants:

11. What are your plans for the **storage of data** (electronic, digital, paper, etc.)? Please ensure that your plans comply with the General Data Protection Regulation (GDPR) and the (UK) Data Protection Act 2018.

12. Please answer the following question *only* if you have not completed a Data Management Plan for an external funder.

12.a How will you make your data available under open access requirements?

12b. Are there any restrictions on sharing your data for open access purposes?

13. Will **audio or video recording** take place? ☐ no ☐ audio ☐ video

13a. Please confirm that portable devices (laptop, USB drive etc) will be **encrypted** where they are used for identifiable data. If it is not possible to encrypt your portable devices, please comment on the steps you will take to protect the data.

13b. What arrangements have been made for **audio/video data storage**? At what point in the research will tapes/digital recordings/files be destroyed?

13c. If your study includes video recordings, what are the implications for participants' anonymity? Can anonymity be guaranteed and if so, how? If participants are identifiable on the recordings, how will you explain to them what you will do with the recordings? How will you seek consent from them?

14. What are the plans for dissemination of findings from the research? If you are a student, mention here your thesis. Please also include any impact activities and potential ethical issues these may raise.

15. What particular ethical considerations, not previously noted on this application, do you think there are in the proposed study? Are there any matters about which you wish to seek guidance from the FSTREC?

Section Five

Additional information required by the university insurers

If the research involves either the nuclear industry or an aircraft or the aircraft industry (other than for transport), please provide details below:

Section Six

Declaration and Signatures

I understand that as Principal Investigator/researcher/PhD candidate I have overall responsibility for the ethical management of the project and confirm the following:

- I have read the Code of Practice, Research Ethics at Lancaster: a code of practice and I am willing to abide by it in relation to the current proposal.
- I will manage the project in an ethically appropriate manner according to: (a) the subject matter involved and (b) the Code of Practice and Procedures of the University.
- On behalf of the University I accept responsibility for the project in relation to promoting good research practice and the prevention of misconduct (including plagiarism and fabrication or misrepresentation of results).
- On behalf of the University I accept responsibility for the project in relation to the observance of the rules for the exploitation of intellectual property.
- If applicable, I will give all staff and students involved in the project guidance on the good practice and ethical standards expected in the project in accordance with the University Code of Practice. (Online Research Integrity training is available for staff and students here.)
- If applicable, I will take steps to ensure that no students or staff involved in the project will be exposed to inappropriate situations.
- I confirm that I have completed all risk assessments and other Health and Safety requirements as advised by my departmental Safety Officer.

☑ Confirmed

Please note: If you are not able to confirm the statement above please contact the FST Research Ethics Committee and provide an explanation.

Student applicants:

Please tick to confirm that you have discussed this application with your supervisor, and that they agree to the application being submitted for ethical review ☒

Students must submit this application from your Lancaster University email address, and copy your supervisor in to the email in which you submit this application

<u>All Staff and Research Students must complete this declaration:</u>
I confirm that I have sent a copy of this application to my Head of Department (or their delegated representative) . **Tick here to confirm** ☒
Name of Head of Department *(or their delegated representative)*
Adrian Friday / Mark Rouncefield

Applicant electronic signature: Lama Date 11 April 2019

A.2 No. of tweets related to Infectious Diseases

Here is a table showing the number of tweets related to Infectious Diseases in 2019, 2020, and 2021.

المرض المعدي	Infectious Diseases	Sep-19-12	Sep-19-23	Oct-19-01	Oct-19-09	Oct-19-16	Oct-19-26	Nov-19-06	Nov-19-14	Nov-19-25	Dec 3 2019	Dec-19-10	Dec-19-17
الحمى الشوكية (التهاب السحايا)	Meningitis	200	131	40	79	31	99	766	220	40	107	264	73
الحصبة	Measles	300	463	404	506	256	190	960	638	488	1495	1397	1144
الانفلونزا الموسمية	Influenza	3391	3391	3045	4069	2597	7971	10062	7652	9082	7852	16247	4853
الإيدز (نقص المناعة المكتسب)	The Human Immunodeficiency Virus (AIDS)	1249	1249	2440	8441	2606	1533	1148	2381	3800	5766	5195	2747
البلهارسيا	Bilharzia Schistoso) (miasis	227	227	118	66	100	121	91	64	81	75	63	69
الحصبة الألمانية	German Measles (Rubella)	152	152	37	96	51	10	104	23	16	21	17	137
الحمى الصفراء	Yellow Fever	122	122	35	22	38	10	431	5979	651	130	540	234
الحمى المالطية	Malta Fever	13	13	4	5	24	19	8	29	13	18	12	29
الدرن	Tuberculosis	124	124	155	213	373	362	538	235	269	480	238	115
السُل	Tuberculosis	1561	1561	695	736	883	606	1395	1009	667	1247	1380	1668
الطاعون	Plague	1134	1134	1390	1385	1653	1448	974	1633	1313	2144	1895	1218
القمل	Lice	1175	1175	901	1084	1680	1650	823	872	993	929	799	834
الليشمانيا	Leishmaniasis	68	68	25	55	14	7	47	17	107	24	31	10
الملاريا	Malaria	2597	2597	1427	3167	564	2680	2952	1844	1803	3947	3959	1175
النكاف	Mumps	122	122	21	18	28	15	59	12	13	6	26	22
أنفلونزا الطيور	Bird Flu	92	92	48	151	97	51	63	96	100	68	93	162
حمى الضنك	Dengue Fever	178	178	183	315	1608	1575	443	3544	5667	1060	668	4063
داء الفيل	Elephantiasis	177	177	7	1138	442	19	10	205	271	27	20	7
داء الكلب	Rabies	175	175	318	177	166	51	126	299	72	63	73	101
عدوى فيروس الورم الحليمي البشري	Human Papillomavirus Virus (HPV)	23	23	7	13	24	17	60	10	66	30	23	13
فيروس زيكا	Zika Virus	5	5	7	1	4	6	14	31	8	13	8	0
كورونا	Coronavirus	147	147	168	129	125	175	196	272	200	13	280	156

المرض المعدي	Infectious Diseases	Jan-20-06	Jan-20-14	Jan-20-21	Jan-20-28	Feb-20-04	Feb-20-11	Feb-20-18	Feb-20-25	Mar-20-03	Mar-20-10	Mar-20-17	Mar-20-24	Mar-20-31	Apr-20-08	Apr-20-13
الحمى الشوكية (التهاب السحايا)	Meningitis	24	30	24	37	81	85	70	35	73	95	64	20	49	109	87
الحصبة	Measles	285	272	608	250	247	253	378	437	536	1059	1214	1884	1136	688	518
الانفلونزا الموسمية	Influenza	2181	2502	2500	8678	9286	5764	3904	6393	18326	25945	30540	23258	34992	16007	8835
الإيدز (نقص المناعة المكتسب)	The Human Immunodeficiency Virus (AIDS)	1234	1074	753	1528	4221	3791	1383	2868	6533	5663	5730	7964	4020	6686	2897
البلهارسيا	Bilharzia Schistoso) (miasis	29	70	45	71	80	87	84	108	211	133	206	241	180	180	99
الحصبة الألمانية	German Measles (Rubella)	15	24	19	4	3	7	43	27	38	256	181	159	399	29	32
الحمى الصفراء	Yellow Fever	57	39	15	67	59	64	49	21	52	59	108	91	101	193	48
الحمى المالطية	Malta Fever	11	16	46	45	5	11	27	24	10	13	21	111	35	25	21
الدرن	Tuberculosis	117	245	136	355	282	127	155	207	186	231	437	452	3539	1247	878
السُل	Tuberculosis	764	1316	1379	758	701	416	655	821	804	1691	2474	1758	3435	2755	2119
الطاعون	Plague	968	1805	803	5560	7500	4696	3275	12310	34237	37528	55339	102036	46364	27911	32505
القمل	Lice	1005	870	1162	1263	688	1110	2435	2624	1303	2222	1300	1594	930	2571	1830
الليشمانيا	Leishmaniasis	2	5	3	16	16	8	11	12	13	12	3	7	15	6	1
الملاريا	Malaria	482	548	674	1936	550	320	383	1946	3064	3085	5456	44337	13864	6042	5884
النكاف	Mumps	4	10	8	5	2	4	55	60	11	23	9	27	354	44	49
أنفلونزا الطيور	Bird Flu	108	55	110	984	7333	8058	803	1013	2664	2357	46082	33235	1787	1410	4209
حمى الضنك	Dengue Fever	420	766	1038	674	756	556	209	313	836	1218	630	422	587	637	897
داء الفيل	Elephantiasis	27	14	11	16	286	214	20	27	23	13	16	17	13	7	10
داء الكلب	Rabies	76	79	176	194	270	238	115	153	193	279	237	153	107	92	59
عدوى فيروس الورم الحليمي البشري	Human Papillomavirus Virus (HPV)	8	37	23	71	255	70	59	419	67	31	30	24	9	4	14
فيروس زيكا	Zika Virus	1	3	6	36	42	20	12	7	27	39	113	122	42	45	35
كورونا	Coronavirus	256	830	3648	105531	85592	123558	58489	177841	204037	251049	299332	413689	210684	474585	350058

المرض المعدي	Infectious Diseases	Apr-20-20	Apr-20-27	May-20-05	May-20-12	May-20-20	May-20-27	Jun-20-02	Jun-20-09	Jun-20-22	Jul-20-01	Jul-20-08	Jul-20-17	Jul-20-24	Aug-20-01	Sep-20-10
الحمى الشوكية (التهاب السحايا)	Meningitis	24	325	79	61	60	67	24	25	139	51	26	59	113	36	116
الحصبة	Measles	730	513	489	683	587	766	934	544	360	417	387	510	398	336	234
الانفلونزا الموسمية	Influenza	5665	4739	8015	4416	9812	6300	12549	7194	8715	6265	6022	3469	3318	2836	4897
الإيدز (نقص المناعة المكتسب)	The Human Immunodeficiency Virus (AIDS)	7616	2808	5540	3752	6368	2905	3815	3888	1792	2734	2046	939	1789	1785	793
البلهارسيا	Bilharzia Schistoso) (miasis	67	80	56	74	79	598	162	227	100	64	62	68	86	60	45
الحصبة الألمانية	German Measles (Rubella)	65	216	64	27	32	33	12	63	53	15	23	30	50	139	15
الحمى الصفراء	Yellow Fever	90	54	48	18	18	14	18	23	31	10	17	66	16	12	36
الحمى المالطية	Malta Fever	11	10	23	160	18	16	26	32	29	15	20	69	14	9	4
الدرن	Tuberculosis	736	471	454	1314	270	419	545	310	282	182	243	247	353	259	288
السُل	Tuberculosis	2963	947	508	654	990	911	990	744	691	649	630	1208	545	542	904
الطاعون	Plague	11602	13519	6066	7264	10146	3738	8581	7056	3609	2695			2235	2148	1450
القمل	Lice	1005	1017	2070	1957	1184	997	1535	1949	1432	1079	990	1251	1076	750	789
الليشمانيا	Leishmaniasis	0	3	4	3	20	1	0	1	23	4	4	77	83	10	2
الملاريا	Malaria	2616	3835	2381	2106	2215	3188	2863	1927	1303	622	569	397	305	394	755
النكاف	Mumps	21	199	47	36	27	18	14	102	32	12	20	23	31	7	12
أنفلونزا الطيور	Bird Flu	775	642	386	372	1032	387	614	350	679	228	303	190	77	99	70
حمى الضنك	Dengue Fever	266	166	1452	687	916	421	424	221	137	96	197	333	83	88	295
داء الفيل	Elephantiasis	33	22	14	23	5	4	81	21	22	120	21	16	11	14	14
داء الكلب	Rabies	78	218	87	72	70	101	58	80	109	71	179	105	44	79	85
عدوى فيروس الورم الحليمي البشري	Human Papillomavirus Virus (HPV)	22	19	29	33	1	39	8	36	27	31	8	677	350	45	21
فيروس زيكا	Zika Virus	43	13	34	39	111	50	33	59	17	15	7	5	12	2	2
كورونا	Coronavirus	1004211	66466	641617	619622	447078	434366	482876	698289	348326	636346	636347	636348	636349	636350	636351

المرض المعدي	Infectious Diseases	Sep-20-16	Sep-20-23	Sep-20-29	Oct-20-06	Oct-20-13	Oct-20-19	Oct-20-26	Nov-20-03	Nov-20-10	Nov-20-17	Nov-20-24	Dec-20-01	Dec-20-09	Dec-20-16	Dec-20-28
الحمى الشوكية (التهاب السحايا)	Meningitis	148	71	88	137	129	35	44	66	21	53	29	41	37	23	28
الحصبة	Measles	461	557	333	296	336	696	700	359	518	431	342	295	292	421	978
الانفلونزا الموسمية	Influenza	4421	7057	9741	12388	12486	17137	26120	26725	13830	9387	5483	7031	8472	3828	5751
الإيدز (نقص المناعة المكتسب)	The Human Immunodeficiency Virus (AIDS)	965	1193	834	1008	1740	948	788	565	545	601	865	3097	4906	2068	1367
البلهارسيا	Bilharzia Schistoso)(miasis	47	109	37	60	62	80	85	69	31	42	88	69	33	32	56
الحصبة الألمانية	German Measles (Rubella)	9	8	37	40	106	31	57	100	28	26	40	17	49	28	86
الحمى الصفراء	Yellow Fever	10	14	11	18	27	37	13	8	5	41	55	11	49	14	13
الحمى المالطية	Malta Fever	7	10	9	14	49	39	23	51	19	28	15	22	10	9	9
الدرن	Tuberculosis	136	124	125	122	181	134	164	110	173	174	173	392	135	1126	235
السُل	Tuberculosis	706	647	471	554	710	852	473	417	314	593	458	450	424	1480	521
الطاعون	Plague	1412	1339	1196	1385	1879	1403	2069	1151	1070	1237	1348	1042	1381	1340	2049
القمل	Lice	1000	1182	1307	1060	874	1151	1112	623	564	516	605	487	500	1019	646
الليشمانيا	Leishmaniasis	5	1	15	19	23	0	3	4	3	2	32	13	129	1532	24
الملاريا	Malaria	518	747	927	1488	581	699	654	356	305	214	757	1047	594	336	381
النكاف	Mumps	14	6	16	30	21	13	16	15	18	53	113	20	27	19	60
أنفلونزا الطيور	Bird Flu	172	125	101	258	87	75	66	213	457	262	315	494	256	371	513
حمى الضنك	Dengue Fever	1671	187	525	546	173	133	220	149	92	186	317	162	99	103	82
داء الفيل	Elephantiasis	21	26	19	92	44	6	19	16	5	2	12	132	9	11	11
داء الكلب	Rabies	75	41	152	288	134	121	91	36	58	408	75	37	28	39	77
عدوى فيروس الورم الحليمي البشري	Human Papillomavirus Virus (HPV)	9	9	19	40	42	9	8	46	15	10	35	17	29	9	5
فيروس زيكا	Zika Virus	2	1	3	3	0	2	5	4	10	10	3	4	7	16	22
كورونا	Coronavirus	534420	534421	534422	534423	550123	445210	97773	560080	406372	406373	320547	407244	407244	383519	383519

المرض المعدي	Infectious Diseases	Jan-21-05	Jan-21-12	Jan-21-18	Jan-21-25	Feb-21-01	Feb-21-08	Feb-21-15	Feb-21-22	Mar-21-01	Mar-21-08	Mar-21-15	Mar-21-22	Mar-21-29	Apr-21-05	Apr-21-12
الحمى الشوكية (التهاب السحايا)	Meningitis	18	11	80	38	29	151	144	38	27	24	70	29	119	15	18
الحصبة	Measles	381	244	249	225	223	273	255	1694	234	455	298	1206	1267	272	274
الانفلونزا الموسمية	Influenza	2934	3333	5936	3026	3005	4298	2398	2664	2258	2115	2036	2792	2478	2599	2046
الإيدز (نقص المناعة المكتسب)	The Human Immunodeficiency Virus (AIDS)	966	781	713	642	679	1320	1579	843	523	798	2025	808	901	1108	1159
البلهارسيا	Bilharzia Schistoso)(miasis	23	26	59	53	48	31	36	54	67	33	33	119	52	46	61
الحصبة الألمانية	German Measles (Rubella)	107	33	26	28	8	12	20	12	17	77	14	38	135	33	92
الحمى الصفراء	Yellow Fever	20	154	17	5	212	18	6	14	21	14	25	11	10	35	21
الحمى المالطية	Malta Fever	5	18	8	109	8	11	18	13	7	15	24	12	8	8	29
الدرن	Tuberculosis	95	59	94	178	212	198	62	458	435	128	111	257	727	325	229
السُل	Tuberculosis	521	606	538	465	388	348	507	606	402	603	489	658	1811	609	427
الطاعون	Plague	1528	1584	1104	1284	1488	2758	2095	1596	1382	1938	1960	1234	1808	1406	1710
القمل	Lice	826	607	573	587	955	604	595	463	497	550	825	1384	1301	672	715
الليشمانيا	Leishmaniasis	7	1	3	7	10	5	11	1	2	4	24	2	1	15	8
الملاريا	Malaria	400	361	309	155	258	267	288	288	302	318	438	266	401	357	210
النكاف	Mumps	20	11	6	16	13	587	140	41	23	127	20	41	80	32	36
أنفلونزا الطيور	Bird Flu	316	906	661	448	463	570	741	1279	464	122	1519	199	110	104	140
حمى الضنك	Dengue Fever	41	182	114	69	94	54	107	118	75	69	44	82	156	58	64
داء الفيل	Elephantiasis	67	145	25	13	6	10	5	14	5	9	10	3	9	2	3
داء الكلب	Rabies	49	184	211	48	71	41	59	40	104	206	127	180	264	49	63
عدوى فيروس الورم الحليمي البشري	Human Papillomavirus Virus (HPV)	7	30	39	37	46	57	26	17	74	105	26	35	12	8	20
فيروس زيكا	Zika Virus	7	0	5	26	54	21	5	12	54	2	13	23	8	3	0
كورونا	Coronavirus	470338	510562	510563	510564	510565	510566	400852	400853	555345	416443	416443	388767	415316	435107	519371

المرض المعدي	Infectious Diseases	Apr-21-19	Apr-21-26	May-21-03	May-21-17	May-21-24	Jun-21-03	Jun-21-07	Jun-21-14	Jun-21-21	Jun-21-28	Jul-21-05	Jul-21-12	Jul-21-19	Jul-21-26	Aug-21-02
الحمى الشوكية (التهاب السحايا)	Meningitis	114	293	38	13	17	17	43	32	40	42	37	29	117	31	37
الحصبة	Measles	229	326	377	226	625	625	754	677	535	607	518	705	498	247	452
الانفلونزا الموسمية	Influenza	2353	2853	3714	1426	2415	2415	2155	2633	3063	4090	4878	3725	3211	3569	3764
الإيدز (نقص المناعة المكتسب)	The Human Immunodeficiency Virus (AIDS)	995	728	962	792	1169	1169	1443	1753	3954	1397	1474	1442	705	1250	1619
البلهارسيا	Bilharzia Schistoso) (miasis	28	44	58	22	60	60	54	74	39	38	71	79	143	82	35
الحصبة الألمانية	German Measles (Rubella)	51	19	26	14	52	52	43	44	23	45	71	180	28	24	17
الحمى الصفراء	Yellow Fever	5	29	56	6	23	23	42	27	69	38	15	51	41	18	26
الحمى المالطية	Malta Fever	18	9	16	133	13	13	11	221	19	11	8	15	7	106	12
الدرن	Tuberculosis	138	187	226	144	117	117	99	139	203	273	259	868	369	301	177
السُل	Tuberculosis	386	376	389	468	418	418	456	389	490	425	444	772	768	481	625
الطاعون	Plague	1463	2183	2243	1229	1215	1215	1994	1200	1677	1620	1592	1212	1124	1740	1669
القمل	Lice	804	1185	1132	1509	735	735	565	656	510	676	1018	1350	683	634	656
الليشمانيا	Leishmaniasis	2	7	4	0	4	4	2	1	7	2	6	2	2	20	2
الملاريا	Malaria	478	1144	867	399	193	193	210	379	699	754	469	215	269	237	890
النكاف	Mumps	16	6	12	5	28	28	39	44	27	34	11	42	29	16	23
أنفلونزا الطيور	Bird Flu	433	134	67	81	1178	1178	4751	1885	297	349	315	116	311	560	66
حمى الضنك	Dengue Fever	53	58	32	61	42	42	26	76	72	42	23	49	41	37	35
داء الفيل	Elephantiasis	3	28	13	8	5	5	11	8	4	3	11	2	5	4	4
داء الكلب	Rabies	44	23	63	45	98	98	75	65	133	51	43	80	83	40	177
عدوى فيروس الورم الحليمي البشري	Human Papillomavirus Virus (HPV)	7	13	14	8	23	23	61	7	6	7	8	7	7	23	12
فيروس زيكا	Zika Virus	173	3	6	2	10	10	14	4	6	16	0	117	37	6	8
كورونا	Coronavirus	519372	519373	445371	288636	358004	358004	353139	329206	382833	399921	399921	371792	396876	364927	364927

المرض المعدي	Infectious Diseases	Aug-21-09	Aug-21-16	Aug-21-23	Aug-21-30	Sep-21-06	Sep-21-13	Sep-21-20	Sep-21-27
الحمى الشوكية (التهاب السحايا)	Meningitis	24	24	61	20	53	85	143	80
الحصبة	Measles	497	319	374	538	314	460	1207	774
الانفلونزا الموسمية	Influenza	3560	2214	1877	1464	1927	3031	3269	5075
الإيدز (نقص المناعة المكتسب)	The Human Immunodeficiency Virus (AIDS)	863	1091	666	919	705	765	833	724
البلهارسيا	Bilharzia Schistoso)(miasis	39	84	37	39	215	79	26	29
الحصبة الألمانية	German Measles (Rubella)	25	161	32	267	51	21	19	25
الحمى الصفراء	Yellow Fever	25	17	8	27	32	9	8	35
الحمى المالطية	Malta Fever	40	9	13	17	6	18	6	13
الدرن	Tuberculosis	226	161	129	69	109	162	148	245
السُل	Tuberculosis	737	427	386	466	511	398	409	449
الطاعون	Plague	1240	1160	1236	1004	790	1178	1098	1136
القمل	Lice	1024	992	570	574	611	477	723	541
الليشمانيا	Leishmaniasis	4	7	10	2	4	1	3	3
الملاريا	Malaria	903	543	512	293	220	401	182	1735
النكاف	Mumps	16	11	8	28	33	4	5	33
أنفلونزا الطيور	Bird Flu	382	137	99	73	58	130	59	342
حمى الضنك	Dengue Fever	37	43	108	257	109	48	105	55
داء الفيل	Elephantiasis	8	1	8	5	4	12	3	5
داء الكلب	Rabies	251	106	42	70	108	41	29	63
عدوى فيروس الورم الحليمي البشري	Human Papillomavirus Virus (HPV)	18	21	10	4	11	3	3	9
فيروس زيكا	Zika Virus	16	0	5	1	2	0	7	7
كورونا	Coronavirus	364927	305475	283420	351378	278392	256054	228183	35905

Appendix B

An Arabic Infectious Disease Ontology

B.1 Questionnaire Template

Here is the questionnaire template that has been sent to the five different types of people: academic, health professional, students, non-educational people and others to find the slang terms related to infectious diseases.

General terms of infectious diseases

To achieve the objectives of the study:

"Using Twitter to support the analysis of the spread of infectious diseases in Saudi Arabia"

We need to know the informal terms of infectious diseases that might be used on social media to help us find mentions of these diseases. I hope you will fill in the attached questionnaire. Your answers will have a significant impact on the results of the study. All data in this questionnaire will be treated strictly and will only be used for research purposes.

You can add or suggest any name you think fits the name of the disease.

Thank you for your cooperation and good response.

Researcher: Lama Alsudias

Lancaster University

Supervisor: Dr. Paul Rayson

* Required

Age *

- Under 18
- Between 18 and 25
- Between 25 and 35
- Between 35 and 50
- Greater than 50

Sex *

- Male
- female

Do you use Twitter in general *

- Yes
- No

What is the source you are involved in to take information about the disease if it is spread?

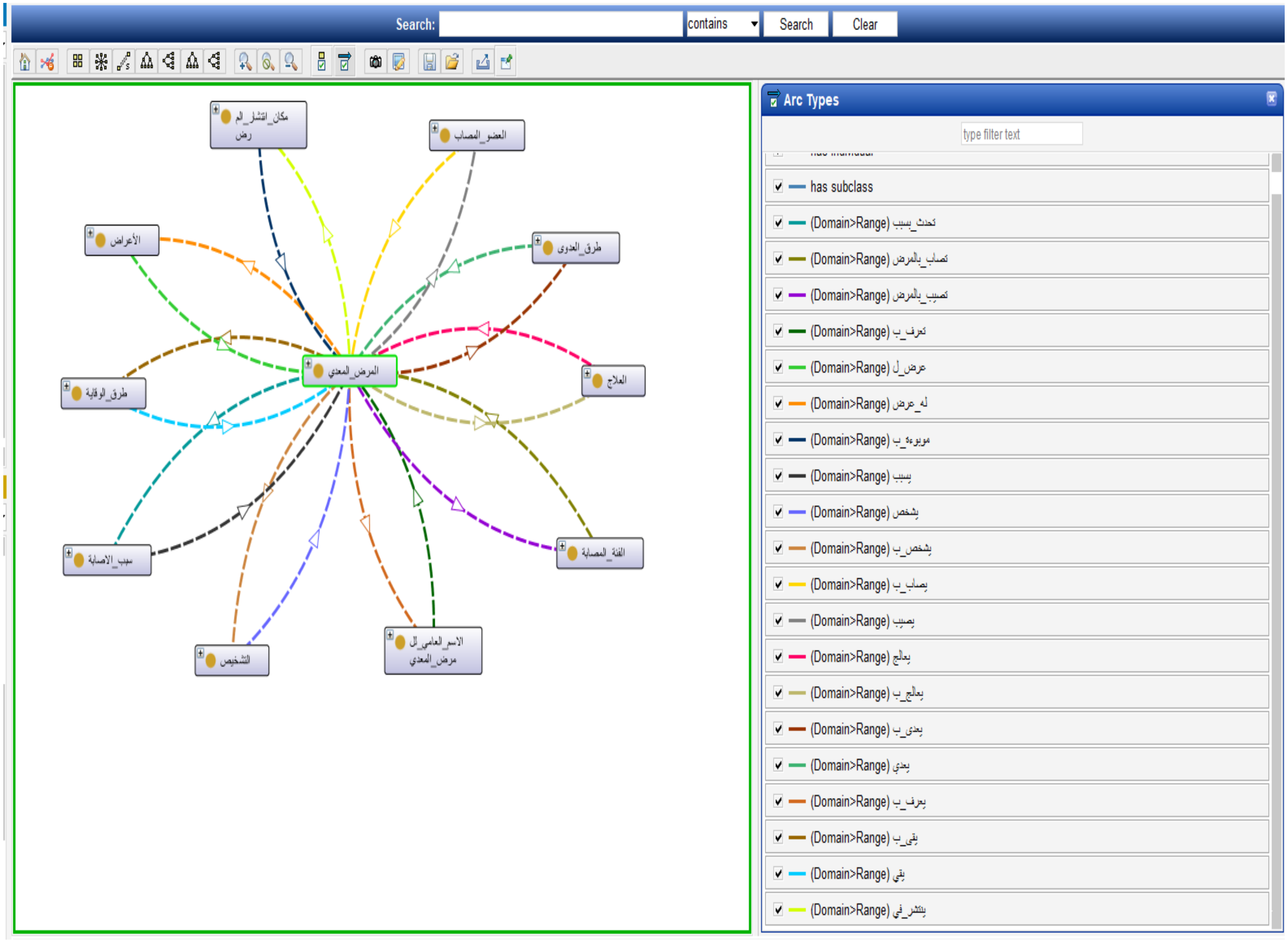
- Website of the Ministry of Health
- TV channels
- newspapers and magazines
- Social media (Twitter)
- Social Networking (WhatsApp)
- Others

What is the general term for (short Answer):

1. meningitis
2. Measles
3. Leishmania
4. Dengue fever
5. Schistosomiasis
6. AIDS (Acquired Immunodeficiency Syndrome)
7. Tuberculosis
8. Vector-borne diseases
9. Immunization
10. Acquired Immunodeficiency Syndrome (HIV / AIDS)
11. Malaria
12. Seasonal flu
13. Avian influenza
14. Maltese fever
15. Yellow fever
16. Rabies
17. Zika virus
18. Plague
19. Tuberculosis
20. Mumps
21. Nipah virus
22. Lice
23. German measles
24. Methicillin-resistant Staphylococcus aureus (Marsa)
25. Elephant disease
26. HPV infection

B.2 Additional Figures

Some extra figures



☑ — has subclass
☑ — (Domain>Range) تحدث_بسبب
☑ — (Domain>Range) تصاب_بالمرض
☑ — (Domain>Range) تصيب_بالمرض
☑ — (Domain>Range) تعرف_ب
☑ — (Domain>Range) مرض_ل
☑ — (Domain>Range) له_مرض
☑ — (Domain>Range) موبوءة_ب
☑ — (Domain>Range) يسبب
☑ — (Domain>Range) يشخص
☑ — (Domain>Range) يشخص_ب
☑ — (Domain>Range) يصاب_ب
☑ — (Domain>Range) يصيب
☑ — (Domain>Range) يعالج
☑ — (Domain>Range) يعالج_ب
☑ — (Domain>Range) يعدى_ب
☑ — (Domain>Range) يعدي
☑ — (Domain>Range) يعرف_ب
☑ — (Domain>Range) يقى_ب
☑ — (Domain>Range) يقي
☑ — (Domain>Range) ينتشر_في

Figure B.1: A sub graph of the developed ontology with classes and their relationships

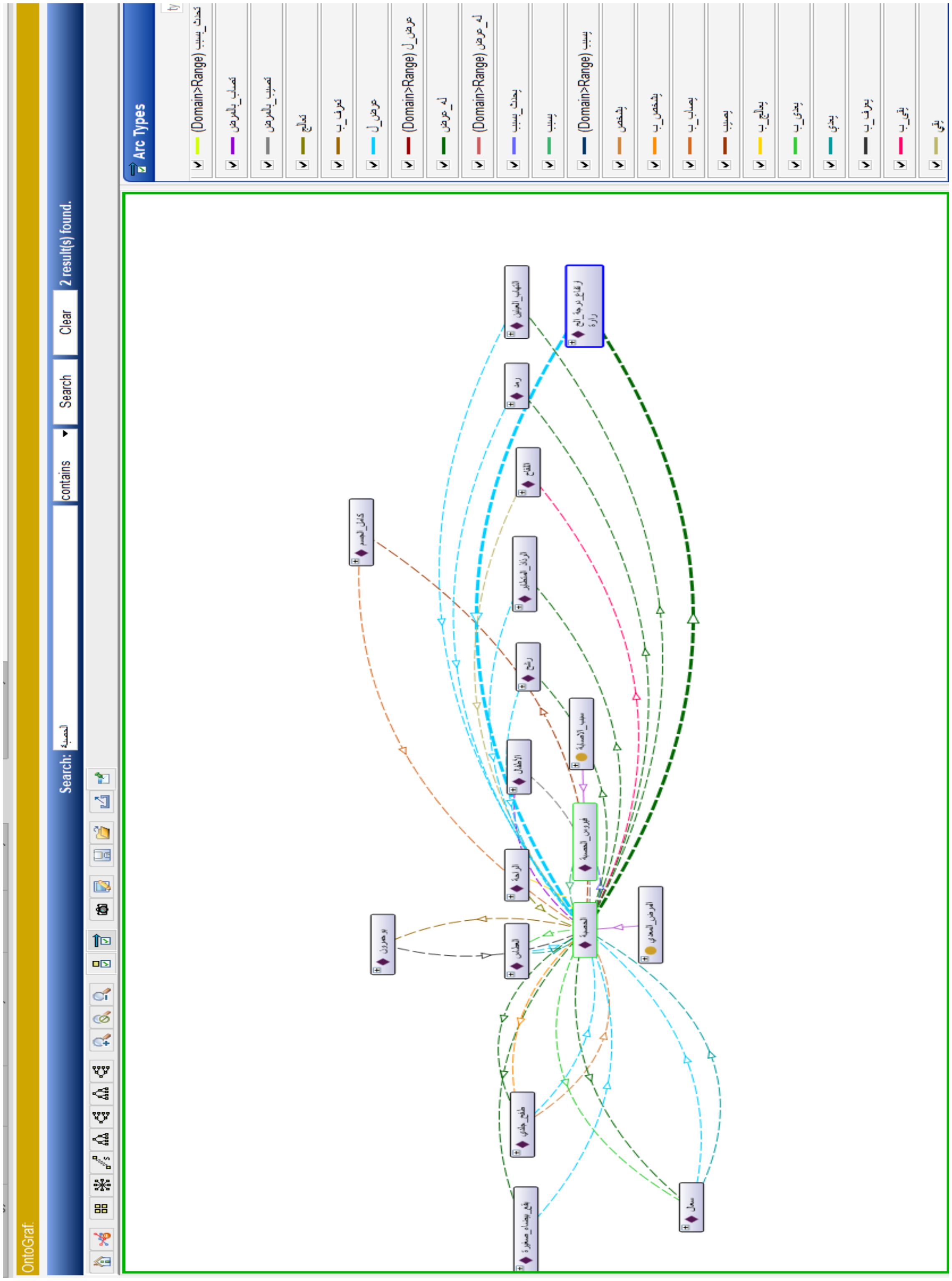


Figure B.2: A screen shot for the Measles disease individuals and relationships

Appendix C

Monitoring Infectious Diseases

C.1 Guidelines for Annotating Infectious Disease Tweets

Here are the guidelines for annotating Infectious Disease tweets, Influenza and COVID-19, that are used in Chapter 7.

Guidelines for Annotating Infectious Disease Tweets

The process is to label them in order to monitor the spread of Influenza and COVID-19. There are 12 classes, 11 originating from the Arabic Infectious Disease Ontology which are: Infectious disease name, i.e. Influenza or COVID-19 in our case, Slang term of infectious disease, Symptom, Cause, Prevention, Infection, Organ, Treatment, Diagnosis, Place of the disease spread, and Infected category and one new class which is 'Infected with', which represents if the user is infected with Influenza or COVID-19. Each tweet is labelled with 1 if the class is included in it or 0 if not included. Moreover, the tweet can be labelled multiple times if it contains multiple classes. For each class described below there is a small description with an example that belong to Influenza case study.

Influenza name: The tweet is labelled with 1 if it includes the Influenza name in it e.g.
حملة تطعيم الانفلونزا تبدأ اليوم
in English, "The flu vaccination campaign starts today".

Slang term of Influenza: The tweet is labelled with 1 if it contains one of the slang terms of Influenza which are:
فلو، النزلة، فلونزا، البرد
e.g.
الفلونزا لاعبة فيني لعب
in English, "Flu is playing with me".

Symptom: The tweet is labelled with 1 when it refers to any Symptom of Influenza which are high fever, headache, cough, runny nose, muscle pain e.g.
مو قادرة اتحرك من قوة الحرارة
in English, "Can not move because high fever".

Cause: The tweet is labelled with 1 if it mentions the virus which is the cause of Influenza e.g.
فيروس الانفلونزا وطرق الوقاية منه
in English, "Influenza virus and ways to prevent it".

Prevention: The tweet is labelled with 1 when it contains one of the preventative methods which are vaccine, washing hands, and cleaning surfaces

e.g.
وزارة الصحة توصي بأخذ لقاح الانفلونزا بدءا من اكتوبر
in English, "The Ministry of Health recommends taking the flu vaccine starting from October".

Infection: The tweet is labelled with 1 if it contains any ways of infection by Influenza such as cough, sneezing, and touching contaminated surfaces e.g.
الطريقة الصحيحة لمنع انتشار رذاذ العطاس بالصور
in English, "The correct way to prevent the spread of sneeze with pictures".

Organ: The tweet is labelled with 1 when it mentions any part of the body affected by Influenza such as head or nose e.g.
خشمي صار طماطة من الانفلونزا
in English, "My nose became tomato from the flu".

Treatment: The tweet is labelled with 1 if it contains one of Influenza treatments which are to drink fluids, rest and take sedatives e.g.
شرب السوائل يساعد في سرعة الشفاء من الانفلونزا
in English, "Drinking fluids helps speed recovery from the flu".

Diagnosis: The tweet is labelled with 1 if it includes the method of diagnosis Influenza which is clinical examination e.g.
ساعة ونص انتظار عند الدكتور وبالنهاية بس بنادول. اكره الانفلونزا وقت الاختبارات
in English, "An hour and a half waiting at the doctor, and at the end, just Panadol. I hate flu in tests time".

Place of the disease spread: The tweet is labelled with 1 if it contains any location e.g.
اكاد اجزم انه مكة كلها موبوءة بالفلونزا
in English, "I can almost confirm that all of regions of Mecca is infected with influenza".

Infected category: The tweet is labelled with 1 if it mentions a category such as kids, old people, or pregnant women e.g.
ماذا تفعل الحامل عند الاصابة بالانفلونزا
in English, "What do pregnant women do when infected with the flu".

Infected with: The tweet is labelled with 1 if it's understood that the user is infected with Influenza e.g.
صداع شديد و الانفلونزا تعبتني
in English, "Severe headache and the flu bored me".

References

Abdul-Mageed, Muhammad et al. (2020). “NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task”. In: *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pp. 97–110.

Abuleil, Saleem, Khalid Alsamara, and Martha Evens (2002). “Acquisition system for Arabic noun morphology”. In: *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*.

Addawood, Aseel et al. (Dec. 2020). “Tracking And Understanding Public Reaction During COVID-19: Saudi Arabia As A Use Case”. In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online: Association for Computational Linguistics. URL: `https://www.aclweb.org/anthology/2020.nlpcovid19-2.24`.

Ahmed, Wasim, Peter Bath, et al. (2018). “Moral panic through the lens of Twitter: An analysis of infectious disease outbreaks”. In: *Proceedings of the 9th International Conference on Social Media and Society*. ACM, pp. 217–221.

Ahmed, Wasim, Peter A Bath, and Gianluca Demartini (2017). “Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges”. In: *The Ethics of Online Research*. Emerald Publishing Limited, pp. 79–107.

Ahmed, Wasim, Peter A Bath, Laura Sbaffi, et al. (2019). “Novel insights into views towards H1N1 during the 2009 Pandemic: a thematic analysis of Twitter data”. In: *Health Information & Libraries Journal* 36.1, pp. 60–72.

Ahmed, Wasim, Gianluca Demaerini, and Peter A Bath (2017). “Topics discussed on twitter at the beginning of the 2014 Ebola epidemic in United States”. In: *iConference 2017 Proceedings*.

Ahmed, Wasim, G. Demartini, and P. Bath (Mar. 2017). “Topics Discussed on Twitter at the Beginning of the 2014 Ebola Epidemic in United States”. In:

Al Ameed, Hayder et al. (2005). “Arabic light stemmer: A new enhanced approach”. In: *The Second International Conference on Innovations in Information Technology (IIT’05)*. Citeseer, pp. 1–9.

AlAgha, Iyad M and Alaa Abu-Taha (2015). “AR2SPARQL: an arabic natural language interface for the semantic web”. In: *AR2SPARQL: An Arabic Natural Language Interface for the Semantic Web* 124.18.

Alanazi, Eisa et al. (2020). “Identifying and Ranking Common COVID-19 Symptoms From Tweets in Arabic: Content Analysis”. In: *Journal of Medical Internet Research* 22.11, e21329.

Alayba, Abdulaziz et al. (Apr. 2017). “Arabic language sentiment analysis on health services”. In: *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pp. 114–118. DOI: 10.1109/ASAR.2017.8067771.

Albadi, Nuha, Maram Kurdi, and Shivakant Mishra (2019). “Hateful people or hateful bots? Detection and characterization of bots spreading religious hatred in Arabic social media”. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW, pp. 1–25.

Albalawi, Yahya, Nikola S Nikolov, and Jim Buckley (Oct. 2019). “Trustworthy Health-Related Tweets on Social Media in Saudi Arabia: Tweet Metadata Analysis”. In: *J Med Internet Res* 21.10, e14731. ISSN: 1438-8871. DOI: 10.2196/14731. URL: http://www.ncbi.nlm.nih.gov/pubmed/31596242.

Albukhitan, Saeed and Tarek Helmy (2013). “Automatic ontology-based annotation of food, nutrition and health Arabic web content”. In: *Procedia Computer Science* 19, pp. 461–469.

Alhanjouri, Mohammed (2017). “Pre Processing Techniques for Arabic Documents Clustering”. In: *International Journal of Engineering and Management Research (IJEMR)* 7.2, pp. 70–79.

Ali, Brahim Ait Ben et al. (2020). “A Recent Survey of Arabic Named Entity Recognition on Social Media”. In: *Revue d'Intelligence Artificielle* 34.2, pp. 125–135.

Alnemer, Khalid et al. (2015). “Are health-related tweets evidence based? Review and analysis of health-related tweets on twitter”. In: *Journal of medical Internet research* 17.10.

Alqurashi, Sarah, Abdulaziz Alashaikh, and Eisa Alanazi (2020). “Identifying Information Superspreaders of COVID-19 from Arabic Tweets”. In: *Preprints*.

Alqurashi, Sarah, Ahmad Alhindi, and Eisa Alanazi (2020). “Large Arabic twitter dataset on Covid-19”. In: *arXiv preprint arXiv:2004.04315*.

Alrefaie, Mohamed Taher (2017). *Arabic stop words*. https://github.com/mohataher/arabic-stop-words.

Alshutayri, Areej Odah O (2018). “Arabic Dialect Texts Classification”. PhD thesis. University of Leeds.

Alsobayel, Hana (Sept. 2016). “Use of Social Media for Professional Development by Health Care Professionals: A Cross-Sectional Web-Based Survey”. In: *JMIR Med Educ* 2.2, e15. ISSN: 2369-3762. DOI: 10.2196/mededu.6232. URL: http://www.ncbi.nlm.nih.gov/pubmed/27731855.

Alsudias, Lama and Paul Rayson (2019). “Classifying information sources in Arabic twitter to support online monitoring of infectious diseases”. In: *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pp. 22–30.

Alsudias, Lama and Paul Rayson (Apr. 2020a). "Analyzing the Spread of Influenza in Arabic Twitter". English. In: HEALTHCARE TEXT ANALYTICS CONFERENCE; Conference date: 23-04-2020. URL: `http://healtex.org/healtac-2020/`.

— (July 2020b). "COVID-19 and Arabic Twitter: How can Arab World Governments and Public Health Organizations Learn from Social Media?" In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics. URL: `https://www.aclweb.org/anthology/2020.nlpcovid19-acl.16`.

— (May 2020c). "Developing an Arabic Infectious Disease Ontology to Include Non-Standard Terminology". In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4844–4852. URL: `https://www.aclweb.org/anthology/2020.lrec-1.596`.

— (June 2021a). "Detecting COVID-19 Misinformation in Arabic Tweets". English. In: HEALTHCARE TEXT ANALYTICS CONFERENCE 2021; Conference date: 17-06-2021 Through 18-06-2021. URL: `http://healtex.org/healtac-2021/`.

— (Sept. 2021b). "Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study". In: *JMIR Med Inform* 9.9, e27670. ISSN: 2291-9694. DOI: `10.2196/27670`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/34346892`.

Antoun, Wissam, Fady Baly, and Hazem Hajj (2020). "AraBERT: Transformer-based Model for Arabic Language Understanding". In: *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pp. 9–15.

Aramaki, Eiji, Sachiko Maskawa, and Mizuki Morita (2011). "Twitter catches the flu: detecting influenza epidemics using Twitter". In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 1568–1576.

Artstein, Ron and Massimo Poesio (2008). "Inter-coder agreement for computational linguistics". In: *Computational Linguistics* 34.4, pp. 555–596.

Arze, Cesar et al. (Nov. 2011). "Disease Ontology: a backbone for disease semantic integration". In: *Nucleic Acids Research* 40.D1, pp. D940–D946. ISSN: 0305-1048. DOI: `10.1093/nar/gkr972`. eprint: `http://oup.prod.sis.lan/nar/article-pdf/40/D1/D940/9483432/gkr972.pdf`. URL: `https://dx.doi.org/10.1093/nar/gkr972`.

Asim, Muhammad Nabeel et al. (2018). "A survey of ontology learning techniques and applications". In: *Database* 2018.

Al-Awadhi, Mohammad, Suhail Ahmad, and Jamshaid Iqbal (2021). "Current Status and the Epidemiology of Malaria in the Middle East Region and Beyond". In: *Microorganisms* 9.2, p. 338.

Ayodele, Taiwo Oladipupo (2010). “Types of machine learning algorithms”. In: *New advances in machine learning* 3, pp. 19–48.

Banda, Juan M et al. (2020). “A large-scale COVID-19 Twitter chatter dataset for open scientific research–an international collaboration”. In: *arXiv preprint arXiv:2004.03688*.

Benajiba, Yassine and Paolo Rosso (2008). “Arabic named entity recognition using conditional random fields”. In: *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*. Vol. 8. Citeseer, pp. 143–153.

Benajiba, Yassine, Paolo Rosso, and José Miguel BenedíRuiz (2007). “ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy”. In: *Computational Linguistics and Intelligent Text Processing*. Ed. by Alexander Gelbukh. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 143–153. ISBN: 978-3-540-70939-8.

Biadsy, Fadi, Julia Hirschberg, and Nizar Habash (2009). “Spoken Arabic dialect identification using phonotactic modeling”. In: *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*, pp. 53–61.

Black, William et al. (2006). “Introducing the Arabic wordnet project”. In: *Proceedings of the third international WordNet conference*. Citeseer, pp. 295–300.

Bojanowski, Piotr et al. (2017). “Enriching word vectors with subword information”. In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Breland, Jessica et al. (2017). “Social Media as a Tool to Increase the Impact of Public Health Research”. In: *American Journal of Public Health* 107.12. PMID: 29116846, pp. 1890–1891. DOI: `10.2105/AJPH.2017.304098`. eprint: `https://doi.org/10.2105/AJPH.2017.304098`. URL: `https://doi.org/10.2105/AJPH.2017.304098`.

Burnaev, Evgeny, Pavel Erofeev, and Artem Papanov (2015). “Influence of resampling on accuracy of imbalanced classification”. In: *Eighth International Conference on Machine Vision (ICMV 2015)*. Vol. 9875. International Society for Optics and Photonics, p. 987521.

Burton, Scott H et al. (2012). “Right time, right place health communication on Twitter: value and accuracy of location information”. In: *Journal of medical Internet research* 14.6, e156.

Campos, Ricardo et al. (2019). “YAKE! Keyword Extraction from Single Documents using Multiple Local Features”. In: *Information Sciences*.

Chandrasekaran, Ranganathan et al. (2020). “Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study”. In: *Journal of medical Internet research* 22.10, e22624.

Charles-Smith, Lauren et al. (2015). “Using social media for actionable disease surveillance and outbreak management: a systematic literature review”. In: *PloS one* 10.10, e0139701.

Chen, Emily, Kristina Lerman, and Emilio Ferrara (2020). “Covid-19: The first public coronavirus twitter dataset”. In: *arXiv preprint arXiv:2003.07372*.

Cimiano, Philipp (2006). “Ontology learning from text”. In: *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*, pp. 19–34.

Collier, Nigel et al. (2010). “An ontology-driven system for detecting global health events”. In: *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 215–222.

Cowell, Lindsay Grey and Barry Smith (2010). “Infectious disease ontology”. In: *Infectious disease informatics*. Springer, pp. 373–395.

Dai, Xiangfeng, Marwan Bikdash, and Bradley Meyer (2017). “From social media to public health surveillance: Word embedding based clustering method for twitter classification”. In: *SoutheastCon 2017*. IEEE, pp. 1–7.

Darwish, Kareem, Hany Hassan, and Ossama Emam (2005). “Examining the effect of improved context sensitive morphology on Arabic information retrieval”. In: *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pp. 25–30.

Darwish, Kareem, Walid Magdy, and Ahmed Mourad (2012). “Language processing for arabic microblog retrieval”. In: *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2427–2430.

Devlin, Jacob et al. (2018). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805*.

Doan, Son et al. (2018). “Using natural language processing to extract health-related causality from Twitter messages”. In: *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*. IEEE, pp. 84–85.

Dredze, Mark et al. (2013). “Carmen: A twitter geolocation system with applications to public health”. In: *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*. Vol. 23. Citeseer, p. 45.

Elhadad, Mohamed K., Kin Fun Li, and Fayez Gebali (2021). “COVID-19-FAKES: A Twitter (Arabic/English) Dataset for Detecting Misleading Information on COVID-19”. In: *Advances in Intelligent Networking and Collaborative Systems*. Ed. by Leonard Barolli, Kin Fun Li, and Hiroyoshi Miwa. Cham: Springer International Publishing, pp. 256–268. isbn: 978-3-030-57796-4.

Farghaly, Ali and Khaled Shaalan (2009). “Arabic natural language processing: Challenges and solutions”. In: *ACM Transactions on Asian Language Information Processing (TALIP)* 8.4, pp. 1–22.

Farzindar, Atefeh and Diana Inkpen (2017). “Natural language processing for social media”. In: *Synthesis Lectures on Human Language Technologies* 10.2, pp. 1–195.

Frantzi, Katerina, Sophia Ananiadou, and Hideki Mima (2000). “Automatic recognition of multi-word terms: the c-value/nc-value method”. In: *International journal on digital libraries* 3.2, pp. 115–130.

Froud, Hanane et al. (2010). "Stemming and similarity measures for Arabic Documents Clustering". In: *2010 5th International Symposium On I/V Communications and Mobile Network*. IEEE, pp. 1–4.

Fung, I.C.-H et al. (May 2016). "Social Media's Initial Reaction to Information and Misinformation on Ebola, August 2014: Facts and Rumors". In: *Public Health Reports* 131, pp. 461–473. DOI: `10.1177/003335491613100312`.

Fung, Isaac Chun-Hai et al. (Dec. 2016). "Ebola virus disease and social media: A systematic review". In: *American journal of infection control* 44.12, pp. 1660–1671. ISSN: 0196-6553. DOI: `10.1016/j.ajic.2016.05.011`. URL: `https://doi.org/10.1016/j.ajic.2016.05.011`.

Ghenai, Amira and Yelena Mejova (2017). "Catching Zika fever: Application of crowdsourcing and machine learning for tracking health misinformation on Twitter". In: *arXiv preprint arXiv:1707.03778*.

Ghosh, Siddhartha (2009). "Application of Natural Language Processing (NLP) Techniques in E–Governance". In: *E-Government Development and Diffusion: Inhibitors and Facilitators of Digital Democracy*. IGI Global, pp. 122–132.

Green, Rachel M and John W Sheppard (2013). "Comparing frequency-and style-based features for twitter author identification". In: *The Twenty-Sixth International FLAIRS Conference*.

Guidère, Mathieu (2020). "NLP Applied to Online Suicide Intention Detection". In: *HealTAC 2020*.

Habash, Nizar (2010). "Introduction to Arabic natural language processing". In: *Synthesis Lectures on Human Language Technologies* 3.1, pp. 1–187.

Habash, Nizar and Owen Rambow (2005). "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop". In: *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pp. 573–580.

Hadziabdic, Emina and Katarina Hjelm (2014). "Arabic-speaking migrants' experiences of the use of interpreters in healthcare: a qualitative explorative study". In: *International journal for equity in health* 13.1, p. 49.

El-Haj, Mahmoud (2012). "Arabic multi-document text summarisation". PhD thesis. University of Essex.

Hamoui, Btool, Abdulaziz Alashaikh, and Eisa Alanazi (2020). "COVID-19: What Are Arabic Tweeters Talking About?" In: *Computational Data and Social Networks*. Ed. by Sriram Chellappan, Kim-Kwang Raymond Choo, and NhatHai Phan. Cham: Springer International Publishing, pp. 425–436. ISBN: 978-3-030-66046-8.

Haouari, Fatima et al. (2020a). "Arcov-19: The first arabic covid-19 twitter dataset with propagation networks". In: *arXiv preprint arXiv:2004.05861* 3.1.

— (2020b). "The First Arabic COVID-19 Twitter Dataset with Propagation Networks". In: *arXiv preprint arXiv:2004.05861*.

Hayate, ISO, Shoko Wakamiya, and Eiji Aramaki (2016). “Forecasting word model: Twitter-based influenza surveillance and prediction”. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 76–86.

Helwe, Chadi and Shady Elbassuoni (2019). “Arabic named entity recognition via deep co-learning”. In: *Artificial Intelligence Review* 52.1, pp. 197–215.

Hong, Yang and Richard Sinnott (2018). “A Social Media Platform for Infectious Disease Analytics”. In: *International Conference on Computational Science and Its Applications*. Springer, pp. 526–540.

Influenza (2021). URL=https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal). Accessed: [2021-05-05].

Inuwa-Dutse, Isa, Mark Liptrott, and Ioannis Korkontzelos (2018). “Detection of spam-posting accounts on Twitter”. In: *Neurocomputing* 315, pp. 496–511. ISSN: 0925-2312. DOI: `https://doi.org/10.1016/j.neucom.2018.07.044`. URL: `https://www.sciencedirect.com/science/article/pii/S0925231218308798`.

Jain, Vinay Kumar and Shishir Kumar (2015). “An effective approach to track levels of influenza-A (H1N1) pandemic in India using twitter”. In: *Procedia Computer Science* 70, pp. 801–807.

Ji, Xiang, Soon Ae Chun, and James Geller (2012). “Epidemic outbreak and spread detection system based on twitter data”. In: *International Conference on Health Information Science*. Springer, pp. 152–163.

— (2013). “Monitoring public health concerns using twitter sentiment classifications”. In: *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*. IEEE, pp. 335–344.

— (2016). “Knowledge-based tweet classification for disease sentiment monitoring”. In: *Sentiment Analysis and Ontology Engineering*. Springer, pp. 425–454.

Jordan, Sophie E et al. (2019). “Using Twitter for public health surveillance from monitoring and prediction to public response”. In: *Data* 4.1, p. 6.

Joshi, Aditya, Sarvnaz Karimi, Ross Sparks, Cécile Paris, and C Raina Macintyre (2019a). “Survey of Text-based Epidemic Intelligence: A Computational Linguistics Perspective”. In: *ACM Computing Surveys (CSUR)* 52.6, pp. 1–19.

— (Oct. 2019b). “Survey of Text-based Epidemic Intelligence: A Computational Linguistics Perspective”. In: *ACM Comput. Surv.* 52.6, 119:1–119:19. ISSN: 0360-0300. DOI: `10.1145/3361141`. URL: `http://doi.acm.org/10.1145/3361141`.

Kalyanam, Janani et al. (2015). “Facts and Fabrications about Ebola: A Twitter Based Study”. In: *ArXiv* abs/1508.02079.

Kanan, Tarek et al. (2019). “A review of natural language processing and machine learning tools used to analyze arabic social media”. In: *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. IEEE, pp. 622–628.

Kapoor, Bhaskar and Savita Sharma (2010). “A comparative study ontology building tools for semantic web applications”. In: *International Journal of Web & Semantic Technology (IJWesT)* 1.3, pp. 1–13.

Keretna, Sara, Ahmad Hossny, and Doug Creighton (2013). “Recognising user identity in twitter social networks via text mining”. In: *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, pp. 3079–3082.

Khanwalkar, Saurabh et al. (2013). “Content-based geo-location detection for placing tweets pertaining to trending news on map”. In: *The Fourth International Workshop on Mining Ubiquitous and Social Environments*, p. 37.

Khoja, Shereen and Roger Garside (1999). “Stemming arabic text”. In: *Lancaster, UK, Computing Department, Lancaster University.*

Kleinbaum, David G et al. (2002). *Logistic regression.* Springer.

Koh, Jing Xuan and Tau Ming Liew (2020). “How loneliness is talked about in social media during COVID-19 pandemic: Text mining of 4,492 Twitter feeds”. In: *Journal of Psychiatric Research.* ISSN: 0022-3956. DOI: `https://doi.org/10.1016/j.jpsychires.2020.11.015`. URL: `https://www.sciencedirect.com/science/article/pii/S0022395620310748`.

Lamb, Alex, Michael Paul, and Mark Dredze (2013). “Separating fact from fear: Tracking flu infections on twitter”. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 789–795.

Li, T., C. Zhang, and S. Zhu (2006). “Empirical Studies on Multi-label Classification”. In: *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’06)*, pp. 86–92. DOI: `10.1109/ICTAI.2006.55`.

Litin, Scott C and Sanjeev Nanda (2009). *Mayo clinic family health book.* Time Incorporated Home Entertainment.

Lopez, Christian E, Malolan Vasu, and Caleb Gallemore (2020). “Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset”. In: *arXiv preprint arXiv:2003.10359.*

Love, Brad et al. (2013). “Twitter as a source of vaccination information: content drivers and what they are saying”. In: *American journal of infection control* 41.6, pp. 568–570.

Luo, Xiao, Gregory Zimet, and Setu Shah (2019). “A natural language processing framework to analyse the opinions on HPV vaccination reflected in twitter over 10 years (2008-2017)”. In: *Human vaccines & immunotherapeutics* 15.7-8, pp. 1496–1504.

Ma, Edward (2018). *3 basic approaches in Bag of Words which are better than Word Embeddings.* URL: `https://towardsdatascience.com/3-basic-approaches-in-bag-of-words-which-are-better-than-word-embeddings-c2cbc7398016` (visited on 07/22/2018).

Madjarov, Gjorgji et al. (2012). “An extensive experimental comparison of methods for multi-label learning”. In: *Pattern recognition* 45.9, pp. 3084–3104.

Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews (2014). “Home location identification of twitter users”. In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 5.3, pp. 1–21.

Maiya, Arun S. (2020). “ktrain: A Low-Code Library for Augmented Machine Learning”. In: *arXiv* arXiv:2004.10703 [cs.LG].

Manguri, Kamaran H, Rebaz N Ramadhan, and Pshko R Mohammed Amin (2020). “Twitter sentiment analysis on worldwide COVID-19 outbreaks”. In: *Kurdistan Journal of Applied Research*, pp. 54–65.

Manning, Christopher et al. (2014). “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60. URL: `http://www.aclweb.org/anthology/P/P14/P14-5010`.

El-Mawass, Nour and Saad Alaboodi (2016). “Detecting Arabic spammers and content polluters on Twitter”. In: *2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC)*, pp. 53–58. DOI: `10.1109/ICDIPC.2016.7470791`.

Molaei, Soheila et al. (2019). “Predicting the spread of influenza epidemics by analyzing twitter messages”. In: *Health and Technology* 9.4, pp. 517–532.

Mouty, Rabeaa and Achraf Gazdar (2018). “Survey on Steps of Truth Detection on Arabic Tweets”. In: *2018 21st Saudi Computer Society National Computer Conference (NCC)*. IEEE, pp. 1–6.

Mubarak, Hamdy and Sabit Hassan (2020). “ArCorona: Analyzing Arabic Tweets in the Early Days of Coronavirus (COVID-19) Pandemic”. In: *arXiv preprint arXiv:2012.01462*.

Muller, Annegret (2018). “Ensemble methods in multi-label classification”. PhD thesis. Stellenbosch: Stellenbosch University.

Musen, Mark A et al. (2015). “The protégé project: a look back and a look forward”. In: *AI matters* 1.4, p. 4.

Nadri, Hamed et al. (2017). “The top 100 articles in the medical informatics: a bibliometric analysis”. In: *Journal of medical systems* 41.10, pp. 1–12.

Nassif, Ali Bou et al. (2019). “Speech recognition using deep neural networks: A systematic review”. In: *IEEE access* 7, pp. 19143–19165.

National Address Maps (2020). URL=https://maps.address.gov.sa/. Accessed: [2020-09-01].

Obeid, Ossama et al. (May 2020). “CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing”. English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European

Language Resources Association, pp. 7022–7032. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.868.

Odlum, Michelle and Sunmoo Yoon (2015). “What can we learn about the Ebola outbreak from tweets?” In: *American journal of infection control* 43.6, pp. 563–571.

Ortiz-Martınez, Yeimer and Luisa F Jiménez-Arcia (2017). “Yellow fever outbreaks and Twitter: Rumors and misinformation”. In: *American journal of infection control* 45.7, pp. 816–817.

Paul, Michael and Mark Dredze (2011). “You are what you Tweet: Analyzing Twitter for public health.” In: *Icwsm* 20, pp. 265–272.

Paul, Michael, Abeed Sarker, et al. (2016). “Social media mining for public health monitoring and surveillance”. In: *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific, pp. 468–479.

Paul, Michael J and Mark Dredze (2012). “A model for mining public health topics from Twitter”. In: *Health* 11.16-16, p. 1.

Pedregosa, Fabian et al. (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Platt, John C. (1999). “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods”. In: *ADVANCES IN LARGE MARGIN CLASSIFIERS*. MIT Press, pp. 61–74.

Qazi, Umair, Muhammad Imran, and Ferda Ofli (2020). “GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information”. In: *SIGSPATIAL Special* 12.1, pp. 6–15.

Raisan, Miaad and Abdulhussien M Abdullah (2017). *Building a Core Arabic Ontology for Iraqi News*. Noor Publishing.

Al-Safadi, Lilac, Mai Al-Badrani, and Meshael Al-Junidey (2011). “Developing ontology for Arabic blogs retrieval”. In: *International Journal of Computer Applications* 19.4, pp. 40–45.

Salah, Ramzi Esmail and L Qadri binti Zakaria (2017). “A comparative review of machine learning for Arabic named entity recognition”. In: *International Journal on Advanced Science, Engineering and Information Technology* 7.2, pp. 511–518.

Saleh, Layan M Bin and Hend S Al-Khalifa (2009). “AraTation: an Arabic semantic annotation tool”. In: *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*, pp. 447–451.

Sarker, Abeed et al. (2020). “Self-reported COVID-19 symptoms on Twitter: An analysis and a research resource”. In: *medRxiv*.

Saunders-Hastings, Patrick R and Daniel Krewski (2016). “Reviewing the history of pandemic influenza: understanding patterns of emergence and transmission”. In: *Pathogens* 5.4, p. 66.

Sawalha, MS and ES Atwell (2010). “Constructing and using broad-coverage lexical resource for enhancing morphological analysis of Arabic”. In: *Proceedings of the* *seventh conference on international language resources and evaluation (LREC'10)*. European Language Resources Association (ELRA), pp. 282–287.

Schriml, Lynn Marie et al. (2011). "Disease Ontology: a backbone for disease semantic integration". In: *Nucleic acids research* 40.D1, pp. D940–D946.

Schütze, Hinrich, Christopher D Manning, and Prabhakar Raghavan (2008). *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.

Semino, Elena et al. (Mar. 2017). "The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study". English. In: *BMJ Supportive and Palliative Care* 7.1. This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/4.0/, pp. 60–66. ISSN: 2045-435X. DOI: 10.1136/bmjspcare-2014-000785.

Shaalan, Khaled (2014). "A survey of arabic named entity recognition and classification". In: *Computational Linguistics* 40.2, pp. 469–510.

Shaalan, Khaled and Mai Oudah (2014). "A hybrid approach to Arabic named entity recognition". In: *Journal of Information Science* 40.1, pp. 67–87.

Shalev-Shwartz, Shai and Shai Ben-David (2014a). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. DOI: 10.1017/CBO9781107298019.

— (2014b). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shuja, Junaid et al. (2020). "Covid-19 open source data sets: A comprehensive survey". In: *Applied Intelligence*, pp. 1–30.

Sinnenberg, Lauren et al. (2017). "Twitter as a Tool for Health Research: A Systematic Review". In: *American Journal of Public Health* 107.1. PMID: 27854532, e1–e8. DOI: 10.2105/AJPH.2016.303512. eprint: https://doi.org/10.2105/AJPH.2016.303512. URL: https://doi.org/10.2105/AJPH.2016.303512.

Sivasankari, Si, Mu Kavitha, and Gi Saranya (2017). "Medical Analysis and Visualisation of Diseases using Tweet data". In: *Research Journal of Pharmacy and Technology* 10.12, pp. 4306–4312.

Al-Smadi, Mohammad et al. (2020). "Transfer learning for arabic named entity recognition with deep neural networks". In: *IEEE Access* 8, pp. 37736–37745.

Soliman, Taysir Hassan et al. (2014). "Sentiment analysis of Arabic slang comments on facebook". In: *International Journal of Computers & Technology* 12.5, pp. 3470–3478.

Sorower, Mohammad S (2010). "A literature survey on algorithms for multi-label learning". In: *Oregon State University, Corvallis* 18, pp. 1–25.

Su, Yue et al. (2020). "Examining the impact of COVID-19 lockdown in Wuhan and Lombardy: a psycholinguistic analysis on Weibo and Twitter". In: *International journal of environmental research and public health* 17.12, p. 4552.

Suarez-Lledo, Victor and Javier Alvarez-Galvez (Jan. 2021). "Prevalence of Health Misinformation on Social Media: Systematic Review". In: *J Med Internet Res* 23.1, e17187. ISSN: 1438-8871. DOI: `10.2196/17187`. URL: `http://www.ncbi.nlm.nih.gov/pubmed/33470931`.

Sutton, C. and A. McCallum (2012). DOI: `10.1561/2200000013`.

Szymanski, Piotr and Tomasz Kajdanowicz (2019). "Scikit-multilearn: a scikit-based Python environment for performing multi-label classification". In: *The Journal of Machine Learning Research* 20.1, pp. 209–230.

Szymański, Piotr and Tomasz Kajdanowicz (2017). "A scikit-based Python environment for performing multi-label classification". In: *arXiv preprint arXiv:1702.01460*.

Takeuchi, Ryo et al. (2017). "Multivariate Linear Regression of Symptoms-related Tweets for Infectious Gastroenteritis Scale Estimation". In: *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, pp. 18–25.

Al-Thubaity, Abdul Mohsen et al. (2014). "Automatic Arabic term extraction from special domain corpora". In: *2014 International Conference on Asian Language Processing (IALP)*. IEEE, pp. 1–5.

Tornes, Adam and Leanne Trujillo (2021). *Enabling the future of academic research with the Twitter API*. `https://bit.ly/3Ct3j96`. Accessed: 2021-11-18.

Tsoumakas, Grigorios and Ioannis Katakis (2007). "Multi-label classification: An overview". In: *International Journal of Data Warehousing and Mining (IJDWM)* 3.3, pp. 1–13.

Uprety, Sagar, Dimitris Gkoumas, and Dawei Song (2020). "A survey of quantum theory inspired approaches to information retrieval". In: *ACM Computing Surveys (CSUR)* 53.5, pp. 1–39.

Varol, Onur et al. (2017). "Online human-bot interactions: Detection, estimation, and characterization". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 11. 1.

Verspoor, Karin et al., eds. (July 2020). *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics. URL: `https://www.aclweb.org/anthology/2020.nlpcovid19-acl.0`.

Versteegh, Kees (2014). *Arabic language*. Edinburgh University Press.

Vorovchenko, Tatiana et al. (Sept. 2017). "Ebola and Twitter. What Insights Can Global Health Draw from Social Media?" In: pp. 85–98. ISBN: 978-3-319-62988-9. DOI: `10.1007/978-3-319-62990-2_5`.

Wang, Sida I and Christopher D Manning (2012). "Baselines and bigrams: Simple, good sentiment and topic classification". In: *Proceedings of the 50th Annual*

Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 90–94.

Wang, Ye et al. (2017). “Comparisons and selections of features and classifiers for short text classification”. In: *IOP Conference Series: Materials Science and Engineering*. Vol. 261. 1. IOP Publishing, p. 012018.

Wang, Yuxi et al. (2019). “Systematic literature review on the spread of health-related misinformation on social media”. In: *Social science & medicine* 240, p. 112552.

Wu, Xi-Zhu and Zhi-Hua Zhou (2017). “A unified view of multi-label performance measures”. In: *International Conference on Machine Learning*. PMLR, pp. 3780–3788.

Xue, Jia et al. (2020). “Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach”. In: *Journal of medical Internet research* 22.11, e20550.

Yang, Sihui, Zifan Ning, and Yaping Wu (2020). “NLP Based on Twitter Information: A Survey Report”. In: *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*. IEEE, pp. 620–625.

Yepes, Antonio, Andrew MacKinlay, and Bo Han (2015). “Investigating public health surveillance using twitter”. In: *Proceedings of BioNLP 15*, pp. 164–170.

Zirikly, Ayah and Mona Diab (2015). “Named entity recognition for arabic social media”. In: *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pp. 176–185.

Al-Zoghby, Aya M., Aya Elshiwi, and Ahmed Atwan (2018). “Semantic Relations Extraction and Ontology Learning from Arabic Texts—A Survey”. In: *Intelligent Natural Language Processing: Trends and Applications*. Ed. by Khaled Shaalan, Aboul Ella Hassanien, and Fahmy Tolba. Cham: Springer International Publishing, pp. 199–225. ISBN: 978-3-319-67056-0. DOI: 10.1007/978-3-319-67056-0_11. URL: https://doi.org/10.1007/978-3-319-67056-0_11.

Zubiaga, Arkaitz, Ahmet Aker, et al. (Feb. 2018a). “Detection and Resolution of Rumours in Social Media: A Survey”. In: *ACM Comput. Surv.* 51.2. ISSN: 0360-0300. DOI: 10.1145/3161603. URL: https://doi.org/10.1145/3161603.

— (2018b). “Detection and resolution of rumours in social media: A survey”. In: *ACM Computing Surveys (CSUR)* 51.2, pp. 1–36.

Zubiaga, Arkaitz, Damiano Spina, et al. (2015). “Real-time classification of twitter trends”. In: *Journal of the Association for Information Science and Technology* 66.3, pp. 462–473.