Dissertation Type: research

DEPARTMENT OF COMPUTER SCIENCE

Public Sentiment on Social Media in Response to Receiving a Covid Vaccine

Noah Sheldon

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Under the Supervision of
Dr. Ayush Joshi

Monday 13th September 2021

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Noah Sheldon, Monday 13th September 2021

Contents

List of Figures

List of Code Listings

Abstract

A pandemic of covid-19 has claimed almost 4.1 million lives worldwide [1]. UK economy is set to suffer more than £700bn of lost output over the next four years due to Covid-19, exacerbated by the government's mishandling of the emergency health situation and Britain's decision to leave the EU, says top economic thinktank [2]. Hospitals experienced a surge in patients [3]. Health care workers were completely under strain [3]. Researchers and pharmaceutical companies were looking for a cure for the disease as it spreads at an alarming rate. Many companies discovered the vaccines within a short period of time [4]. Trials of vaccines and studies of patients were conducted [4]. From April 2020, Moderna will be available [5]. The vaccines for covid were marketed by 14 companies during the time [4]. Almost 20 % of the population has received the first vaccine [6]. The United Kingdom, Canada, and Israel lead the vaccination race [6]. A first dose of the drug was administered on the 8th of December 2020[8]. Several people have now been immunized with the covid vaccine in the last six months.

Opinion mining, also known as sentiment analysis, is a way of using computers to analyse and interpret people's written opinions, thoughts, and feelings [8]. Natural language processing and text mining are a few of the most active fields of research in recent years [8]. There are two main reasons for its popularity. The theory has several applications since opinion plays a major role in almost every aspect of human behaviour [8]. It is in our best interest to hear the opinions of others whenever we are faced with deciding. Secondly, it offers a vast array of scientific challenges not previously encountered [8]. Previously, there were few opinionated texts available in digital forms, which accounted for part of the lack of study [8]. Thus, it should not be surprising that the advent and growth of the social media on the internet coincide with the inception and rapid growth of this field [8]. Its importance to business and to society has contributed to the spread of research outside of computer science into management and social sciences as well [8].

People's sentiments about the covid vaccine can be determined using Twitter data. The covid vaccine has been the topic of several tweets in the last six months. As part of this process, data are collected, pre-processed, positive and negative tweets are separated, and negative tweets are clustered. To access Twitter data, you must first sign up for Twitter developer access, which grants you API keys and tokens. We can directly extract data from Twitter via Python's tweepy package. An Excel file, a CSV file, or a database is used to store the extracted data. Twitter data is analysed to gain an understanding of Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V.

The main achievements of the project are as follows:

- Collection of Twitter data using tweepy API.
- Visualization of the collected data.
- Pre-processing the collected data.
- Performing topic modelling on pre-processed data.
- Visualizing the results of topic modelling.
- Performing classification of the data.

Notation and Acronyms

TP	:	Topic Modelling
API	:	Application programming interface
SMA	:	Social Media Analytics
SLE	:	Systemic Lupus Erythematosus
BoW	:	Bag of words
STM	:	Structured Topic Modelling
LSA	:	Latent Semantic Analysis
LDA	:	Latent Dirichlet Allocation
FHIR	:	Fast healthcare interoperability resource
GP	:	General practitioner
ICD	:	The International Classication of Diseases
KNN	:	K-nearest neighbours
LR	:	Logistic regression
MSE	:	Mean squared error
NPV	:	Negative predictive value
PPV	:	Positive predictive value
PRC	:	Precision recall curve
R-	:	The event that a test produces a negative result
R+	:	The event that a test produces a positive result
ReLU	:	Rectied linear unit
ROC	:	Receiver operating characteristic curve
SGD	:	Stochastic gradient descent
SBMO	:	Sequential based model optimisation
SQL	:	Structured query language
TNR	:	True negative rate
TPE	:	Tree of Parzen estimators
TPR	:	True positive rate

Acknowledgements

I would like to thank my excellent supervisor for his great domain knowledge and consistent support.

Chapter 1

Contextual Background

1.1 Introduction

The motivation for the project is presented in this chapter, demonstrating its importance as a field of study. Next, we want to analyse the impact this project has on stakeholders and how this project can be helpful. Lastly, the main objectives of the project are identified.

1.2 Motivations

1.2.1 A Pandemic

Throughout history, pandemics have emerged, developed, and ended several times, with varying degrees of success, control, and end because of changes in medical understanding, living conditions, and political contexts [9]. Certain viruses and infections may cause different symptoms and affect different populations, and some populations may be more vulnerable than others [9]. Those are the differences between pandemics. COVID-19 mostly affects people over 65 and those with underlying health conditions, unlike influenza 1918, which caused severe symptoms in young and healthy people [10]. A pandemic outbreak like COVID-19 can be understood better if accurate information is reported and diagnostic testing is available, so health companies and governments can take appropriate steps to curb the spread and offer effective advice to people.

Despite being unaware of a severe disease a year ago, the world now has immunity to it because of a historic scientific effort [11]. The COVID-19 vaccine is being developed and tested rapidly [11]. It is estimated that 50 of these drugs are undergoing various stages of clinical testing, according to the WHO in November [12]. There are several ways to inactivate SARS-CoV-2-from crude chemicals to a newer technology never used in vaccine production. The COVID-19 vaccines developed by Pfizer and Moderna may prevent the disease to a degree of 95%, whereas those manufactured by AstraZeneca and Oxford may prove less effective [12]. Many important questions remain, including how effective vaccines are in preventing disease in the elderly and for how long do vaccines provide protection? There is a possibility that vaccines will not prevent the spread of the disease, as they fail to prevent the spread of many other illnesses [12].

Social media websites are flooded with news about the virus [13]. Because of this, these online platforms are experiencing and expressing different points of view, opinions, and emotions during outbreaks [13]. Researchers and computer scientists use big data to understand people's sentiments about current events, especially those related to the pandemic [13]. This will yield remarkable results when these sentiments are analysed [13]. Researchers can generate better decisions and conclusions using historical data. By increasing the availability of data, researchers can generate better decisions [14]. Social media platforms offer more information than ever before, making these data available as current and reasonable sources. It is interesting to note that these data serve as the basis for opinion mining and sentiment analysis.

1.2.2 Analysing sentiment through social media

Platforms offering various forms of electronic communication have enabled humans to communicate and share ideas, information, knowledge, and data amongst themselves [15]. Social media platforms gain remarkable influence and are considered one of the fastest growing information systems for social applications [16]. People use social media applications and spend a lot of time on these outlets [17] and so they are considered the most important sources of big data. According to [18], Facebook, Twitter, Instagram, and Reddit are some of the most frequently used social media services worldwide. Based on the minutes and hours users spend on these applications, as well as their frequency of use [19] and statistical studies, it can be concluded that these applications affect human behaviour. However, despite the extensive analysis of data provided by these social media platforms, they may have adverse psychological effects on people [20], as well as positive psychological effects on people.

These platforms allow people to express their ideas and opinions freely [21]. The value of these posts and opinions becomes obvious when they become assets. The use of social media networks to glean human emotions and entertainment is very useful to the development of business decisions, government policy and influence over international affairs [22]. Opinion mining and sentiment analysis are becoming increasingly useful. By analysing user behaviour using these social media applications, sentiment analysis can contribute to understanding human emotions [23]. It provides the ability to observe large-scale populations at a low cost [24]. Consequently, sentiment analysis is valuable in identifying the key trends and events occurring in society.

According to Statistics [25], Twitter currently has over 300 million accounts, making it one of the most popular social networks. Twitter provides an incredible opportunity to observe people's opinions and sentiments [26]. When deciding whether to post a positive or negative tweet, determining the sentiment is crucial. The limitation of 140 characters that Twitter has adds another challenge. This limits the length of each tweet, which in turn results in people using language that's not related to language processing. Recently, twitter extended the maximum character count to 280 characters per tweet. Twitter contains small text. The Natural Language Processing systems used today have difficulty extracting the sentiment of a person who uses different words and abbreviations. The polarity of the text has been extracted and mined by some researchers using deep learning techniques [27]. Facebook is referred to as FB, before as B4 and oh my god as OMG. This makes sentiment analysis of short texts like Twitter's posts difficult [28].

1.2.3 Advances in Natural Language Processing

Language that is naturally spoken is the language used in our everyday lives. Research in this field has been around for a while, but the popularity of computer science and programming has caused artificial intelligence related research to increase. How we communicate with each other has been profoundly altered by the Internet. So, we began sending texts, emails, voice messages, to, instead of paper mails and letters. Examples of applications which use natural language processing are Google Translate. Grammarly, Chatbots. etc. With Google Translate, machines understand what you are saying and can translate it word by word to the language you are looking for. This is accomplished while maintaining the same meaning of the original sentence. There is good grammar and word recognition in Grammarly. In the past few years, language processing technology has dramatically improved. Additionally, it provides some suggestions for improving the quality of the article by checking the grammar.

The ability of a computer to comprehend human language is called Natural Language Processing (NLP) [29]. NLP is an approach to reading, deciphering, understanding, and processing human languages in a way that is valuable. Natural Language Understanding and Natural Language Generation are the two main components of Natural Language Processing [29]. An artificial intelligence system capable of understanding natural language is considered natural language understanding. As a result, the system can understand the sentences we speak or write. Many real-world problems can be solved because of its use, including Question-Answering, Query Resolution, Sentiment Analysis, Similarity detection in texts, and Automated Chat Bots [29]. If a system understands natural language, only then can it reply to our responses.

A computerized model that can generate text, audio, or other outputs like human-comprehensible language is Natural Language Generation [29]. By using predefined texts datasets, we create sentences using the model. Text is summarized, queries and questions are answered, machine translations (translations from one language into another) are performed, and responses will be generated. NLP has made significant advances in the last two or three years [29]. Pre-trained models are used to solve the required task after they have been trained on large datasets, and their parameters or weights are adjusted. A transfer learning process involves using models that have been previously trained to solve real-world problems. In addition to text classification, part-of-speech identification, named entity recognition, summarizing text, and answering questions, the pre-trained model is fine-tuned to handle tasks like part-of-speech tagging, question answering, etc [29].

An algorithmic form of NLP called Topic Modelling extracts topics from a corpus of texts using algorithms. In a corpus, topics are themes that occur repeatedly. The probability of each topic in the corpus is related to how strongly the topic is present [30]. Words and phrases used in a particular context are evaluated using a vocabulary, or dictionary [31]. Similarity of topics can be explained by topic models using statistical methods. To examine the textual data, the latent Dirichlet allocation (LDA) model was utilized. The LDA model generates a set of topics with probabilities that each is accompanied by a word when applied to the corpus [32]. NLP topic generation thus takes a non-generative approach since topics are comprised of words or phrases already present in the corpus [32]. Moreover, the approach is entirely statistical. It is not pragmatic, syntactic, or semantic. This is due to the absence of contextual analysis, grammar interpretation, and a focus on the frequency of words in the corpus.

1.3 NLP Techniques: An Intro

Techniques used in this project is illustrated in this section. Even though some authors have taken different approaches.

1.3.1 Steps in Sentiment Analysis

User selects the data source from which a sentiment is to be extracted in this phase. A user, for example, can choose from a variety of online sources, such as Facebook, Twitter, and Reddit [22]. Once the data source is selected, the data collection phase begins. Users begin by using a list of keywords [33] or hash tags (for example, #) [34] to obtain the information they want according to their preferences. There are different types of information (e.g., tweets, posts, news, and texts) [22]. The next is the pre-processing phase. The data is prepared for the next phase by processing the extracted information [35]. Stage three includes feature extraction [36] (grammatical structures and mining characteristics), tokenization [37] (converting text into tokens before converting them into vectors) and cleaning [35] (repeated letter removal, text correction, normalisation, stop word removal and language detection). All pre-processed data are then analysed for their intended uses, such as sentiment analysis [35], polarity identification [35], or frequency analysis [33].

1.3.2 Topic Modelling

In recent years, topic models have been widely used in the field of computer science with a focus on text mining and information retrieval. Researchers across many research fields have shown interest in this model since it was first proposed. There have been successful applications of text mining as well in the fields of computer vision [38], population genetics, and social networks [39]. TM is generally effective in summarizing long documents such as news, articles, and books. As microblogs like Twitter have grown in popularity, the need to analyse short texts has become more relevant. It can be quite difficult to detect topics from short text because short text often contains noisy data [40]. The development of TM methods has been made possible due to the difficulty in identifying topics manually, which is impractical and not scalable due to the size of the data. There are several methods to extract topics from short text [41] and standard long text [42]. Several text analysis methods, such as probabilistic latent semantic analysis (PLSA) [43], latent semantic analysis (LSA) [44], and latent Dirichlet allocation (LDA) [45], provide reliable results. There are currently several TM methods that can't learn from short documents. Additionally, TM approaches to short textual data in OSN platforms have major shortcomings, including slang, data sparsity, grammatical errors, unstructured data, insufficient word co-occurrence data, and non-meaningful and noisy words.

1.4 Objectives

To develop a system which identifies the sentiment of public on covid vaccines a list of objectives was produced. These objectives help us picture out the high-level objectives of the project.

1. Collection of Twitter data using tweepy API.
2. Pre-processing the data.
3. Exploring the data.
4. Performing topic modelling and extracting the topics for text with and without emoji strings.
5. Visualizing the results of topic modelling.
6.

Chapter 2

Technical Background

2.1 Introduction

We will provide a literature review regarding the project execution in this chapter. We start by looking at the social media analytics process (SMA). Data from social media can be extracted through the SMA process. Using Twitter data, we move on to sentiment analysis. Understanding public sentiment has been made possible by these methods. On to topic modelling based on twitter data. We can see what topics are being discussed by looking at this. Words that are more commonly used in connection with a topic make up these topics. After discussing topic modelling, we discuss Twitter as a source of data for covid-19. The visualizations provided insights into the literature.

2.2 Literature Review

2.2.1 Introduction

In this section we will talk about the literature relevant to executing the project.

2.2.2 Social Media Analytics

Information can be generated and utilized by government and private parties, both to benefit from the social media opportunity [46]. The public perception of a phenomenon or a particular event can be determined by social media data [46]. Detailed knowledge of the Internet, social media, databases, data structures, big data analysis, data mining, learning algorithms, and data visualization is required to obtain and analyse social media information [46]. An analysis of social media on a particular topic was performed, as well as the creation of prototype devices that used software to retrieve social media information or retrieve data [46]. By analysing and synthesizing social media data, Social Media Analytics (SMA) produces information that can be used by those in need [46]. There are three steps in the SMA process: Capture, Understand, and Present [46].

Using SMA, data is retrieved, stored, processed, and visualized using various clustering and classification algorithms [47]. For example, communities are detected on social media using the community detection algorithm [48]. Basically, social media is taken as one of the research media that can analyse data, which is one of the uses for it [46]. Social media is considered a potential source of data for research projects [46]. Data from social media presents a few challenges: each site uses a different platform volume, complexity from information, and the data is unstructured [49]. SMA meets the challenge by providing tools and frameworks that allow collection, evaluation, analysis, conclusion, and visualization of data linked to social media [50].

Social media data is used in relation to research activities that require tools and frameworks that have been formulated in ways that enable collecting, evaluating, analysing, interpreting, and presenting data in specific ways [46]. The research addresses the problem of tools and frameworks required for supporting research activities involving social phenomena that make use of social media data [46]. The research will discuss SMA at a micro-blogger site due to the complexity and kinds of platforms used by each medium of social media [46]. With the Internet, we can create and share ideas and stories, share information and links, produce, and share multimedia, and create and share knowledge by working in large groups [46]. Social media [51] refers to these services.

Social media platforms offer two types of services. By making and sharing, digital communications are seen as tools for reshaping networking into a community [46]. Consequently, it can create, manage, edit, comment, tag, and allot any information for inclusion as well as connect, connect, and share it [46]. Twitter is currently one of the most popular social media platforms. Microblogging service Twitter lets users share 140-character messages quickly over a variety of platforms. SMA involves establishing and evaluating tools and frameworks for gathering, evaluating, analysing, and visualizing data of social media [52][50]. According to Gartner Research, social media analytics is a way to analyse, measure, and predict social media interactions, topics, ideas, and contents (Gartner Research, Social analytics). Social Media Analysis is about having an analysis process and synthesizing data of social media to provide useful information for stakeholders. SMA consists of three steps, namely Capture, Understand, and Present [50]. In Figure 2.1, you can see how SMA works.

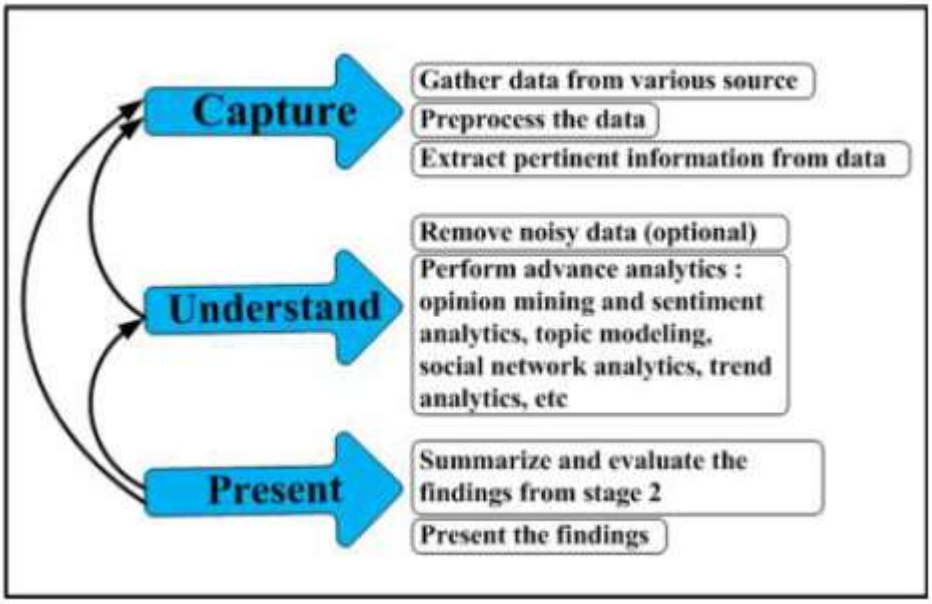

Figure 2.1. Social Media Analytics Process [52]

The capture process of SMA is when we collect data from social media websites [46]. The relevant API/crawler is used to collect data [46]. The data captured is stored in a database. It is then processed to gain information about the data that is relevant to the needs, and the model is then created using the data [52]. Following completion of the Capture process, the next step is to understand the process [46]. It is a process of choosing data relevant to applying data modelling, selecting data containing very few errors, and making a process to analyse the data in such a way that it will provide the best information [52]. Analysing data gets carried out using statistical methods, text mining, data mining, natural language processing (NLP), machine translation, machine learning, and network analysis [52]. It is possible to produce information from social media data using a wide range of techniques, such as opinion mining (or sentiment analysis), topic modelling, social network analysis, trend analysis, and visual analytics [53]. The last step in the process of SMA is present [46]. Initially, present means to visually show or explain, which is a step toward understanding [52]. The information from the analysis process can be visualized in many ways [46].

2.2.3 Sentiment Analysis Using Twitter

Text, speech, tweets, and databases can be used as sources of sentiment analysis as natural language processing (NLP) allows you to extract attitudes, opinions, views, and emotions through an automated process [54]. This is a method of categorizing opinions expressed in the text by marking them as positive, negative, or neutral [54]. This method is also known as opinion mining, appraisal extraction, or subjectivity analysis. There is some overlap in the use of opinion, sentiment, view, and belief, but they differ from one another [54]. Sentiment Analysis encompasses a wide range of tasks like sentiment extraction, sentiment classification, subjectivity classification, opinion summarization, and opinion spam detection, among others [54]. An analysis of emotions, perceptions, opinions, and sentiments toward elements such as products, individuals, topics, organizations, and services [54]. Many researchers have been working on "Sentiment Analysis on Twitter" in recent years [54]. It was originally designed to categorize a review or opinion into two categories, either positive or negative [54].

According to [55], tweets can be classified as objective, positive, or negative. Through the Twitter API, they collected tweets and automatically annotated them using emoticons [55]. In their analysis of the corpus, they constructed a sentiment classifier based on multinomial Naive Bayes, which uses the N-gram and POS-tag features [55]. Due to only having tweets with emoticons in the training set, the set they used was less effective [55]. For tweet classification, [56] implemented two models, the Naive Bayes bigram model, and the Maximum Entropy model. Naive Bayes classifiers outperformed Maximum Entropy classifiers, they found [56]. The proposal presented by [57] was to remove emotional noise from Twitter data using distant supervision, in which the training data was tweeted with emoticons acting as noisy labels. Naive Bayes and Support Vector Machines (SVM) are used to build their models. POS, unigrams, and bigrams made up their feature space. According to their research, conventional models did not outperform SVM, while unigram was more effective as a feature. For classifying tweets, [58] developed an automatic sentiment analysis method consisting of two phases. Second, in phase two, tweets were classified as either positive or negative based on whether they were objective or subjective. There were also features like post polarity and punctuation marks paired with social features such as retweets, hashtags, and links.

Among the methods [59] used was Twitter streaming data. Data from Firehouse API provided all tweets from every user, making them publicly available in real-time [59]. The researchers used multinomial naive Bayes, stochastic gradient descent, and the Hoeffding tree. SGD-based models were found to be the best when used with a learning rate appropriate to the environment [59]. In [60], a three-way model is developed for identifying positive, negative, and neutral sentiment. In addition to unigrams, the researchers also tested feature-based models and tree kernel-based models [60]. Tweets are represented as a tree in kernel-based tree modelling [60]. It is estimated that the feature-based model uses 100 features, and the unigram model uses over 10,000 features [60]. During the evaluation, it was concluded that the features combining prior polarity and the tags attached to parts of speech (POS) are the most important and crucial in the classification process [60]. In comparison with the other two models, the tree kernel-based model performed better [60].

In [61], it was suggested that Twitter users could use hashtags in tweets to identify sentiment by combining punctuation, words, n-grams, and patterns as different traits, then combining them into a common marker vector for sentiment classification. They constructed feature vectors for each of the training and test examples using the K-Nearest Neighbour strategy [61]. Twitter data was collected by [62] using Twitter API. There are three categories of training data (camera, movie, mobile). The data sets are categorized according to the opinions expressed: positive, negative, or non-opinions [62]. Opinion-containing tweets were filtered out [62]. Implementation of the Naive Bayes model and a simplification of independence assumption were undertaken [62]. Additionally, the Chi square and Mutual Information techniques were used to eliminate useless features. Finally, tweets are predicted by their orientation [62]. The outcome could be positive or negative.

Naive Bayes classifiers were presented in [63] to detect polarity in English tweets. Two types of Naive Bayes classifiers were developed: baseline (which was trained to distinguish negative tweets from positive tweets), binary (which used polarity to distinguish positive from negative tweets, ignoring neutral tweets) [63]. In addition to lemmas (nouns, verbs, adjectives, and adverbs), concordance lexicons, and multiword from different sources, classifiers also considered Valence Shifters [63]. An alternative approach to sentiment analysis used [64] based on a bag-of-words strategy in which the relationships between words were not taken into consideration, and documents were viewed as collections of words [64]. The sentiment of every word within the document was determined, and those values were combined with some aggregation functions to determine the sentiment of the whole document [64]. WordNet was used to analyse the emotions associated with words along different dimensions in [65]. To determine semantic polarity, they calculated distance metric on WordNet.

[66] constructed an ensemble framework that combines different feature sets and classification techniques to obtain Sentiment Classification. The authors used two types of features (parts of speech information and word relationships) and three types of classifiers (Naive Bayes, Maximum Entropy, and Support Vector Machines) in their analysis [66]. As a result, they used ensemble modelling approaches like fixed combinations, weighted combinations, and meta-classifiers to improve sentiment classification [66]. There are challenges involved in mining opinions from Twitter tweets and an effective technique to do so was outlined in [67]. In Twitter, opinion retrieval is difficult because of spam and wildly varying language [54].

In the comments or tweets, one can find the sentiment, which can be used for a variety of purposes [70]. In addition, the author [68,69] points out that sentiments can be classified into two groups, namely negative and positive. An expression of sentiment, or sentiment analysis, is a technique of analysing and quantifying a tweet's expressed opinions [71]. Most machine learning algorithms are based on supervised classification approaches that receive binary responses representing positive and negative sentiment [72]. To train classifiers, it is necessary to use labelled data [73]. As a result of this, it is evident that factors such as negative and intensification connotations need to be considered in the evaluation of a word [74]. A strong negative or positive value is negated when it is shifted, which correctly models a mixed perspective. [75] has pointed out that lexical based approaches are less appropriate for Twitter than machine-learning-based approaches. Additionally, machine learning methods can generate an assessment of the most frequently occurring word on Twitter for each integer value assigned based on its frequency of use [70].

An NLP approach uses machine learning, especially statistical learning, which uses an algorithm that combines a large sample of data with a large corpus to learn the rules [76]. Naturally, Language Processing has been used to provide sentiment analysis at various granularities. Classification began as a document-level classification task [77], and later became a sentence-level classification task [78]. An NLP system is a way to make computers interact with the real world by using human language and input to gather meaning [79]. Generally, Support Vector Machines (SVM), Naive Bayes, and N-Gram are the most popular Machine Learning methods [80]. A separating hyperplane defines SVM as a discriminative classifier. A Naive Bayesian classifier is a classifier based on Bayes' theorem and assuming independence between predictors, and it is extremely easy to create with no complicated iterative parameter estimation, which allows it to be used with very large datasets [80]. Furthermore, the Naive Bayes classifier, despite its simplicity, often outperforms more sophisticated classification methods, despite its simplicity [80]. Moreover, N-grams assign probabilities to words or whole sequences within a sentence. Text mining and NLP heavily rely on N-grams.

Speech and language processing utilize this tool extensively. Various tasks have been accomplished using N-grams. Utilizing appropriate feature selection methods, machine learning can remove overlapping and irrelevant features. In contrast, [80] claims that Machine Learning methods suffer greatly from mixed data sets. A high level of classification accuracy is determined by the quality of selected features and the classification algorithm employed. Recent research has focused on finding semantic relationships using word embedding techniques and classification methods based on artificial neural networks [81,82]. As related words usually express the same polarity, the semantic relationship must be examined [83]. Researchers use SVM and Naïve Bayes to compare alternative approaches to their proposed work [83]. Using these two algorithms provides high accuracy with feature selection. [83].

2.2.4 Public Health Research Using social media

Recent years have seen a significant transformation in the way researchers and discoveries are shared, particularly about how results are disseminated [84]. Through the advancement of internet technology, more and more people are interacting, sharing opinions, and debating through social media platforms such as Facebook, Twitter, and Reddit [84]. Forums like these create a community in which people can have interactions and establish relationships [84]. It is possible for these online communities to influence and be influenced by other online communities [84]. Some of the information spread through this spread of influence is known to have affected offline behaviour [85]. As a result of social media content, rumours can spread quickly throughout these communities, and all of this can profoundly impact socioeconomic decisions, political decisions, health care decisions, perceptions, and beliefs as well [86]. It is not new in the public health field to analyse social media texts and to detect social networks [84]. Forecasting clinical surveillance and misinformation have been the subject of many studies [87,88] and [89]. Studies in this area have found a substantial amount of evidence that technologies help the health sector, establishing awareness on social media, and helping people who are living in remote areas [90] or who have limited access to health treatment [91]. These studies generally focus on epidemics and infectious diseases, whereas chronic diseases such as diabetes and cardiovascular disease are the subjects of most effort [92].

The dynamics around online communities have not been fully explored yet, such as the dynamics surrounding systemic lupus erythematosus (SLE), an autoimmune disease whose management is still challenging due primarily to its variety and complexity [84]. SLE patients face many challenges that adversely affect their quality of life and social activities [93]. Aside from that, SLE is also plagued with marked and complex unmet needs [94], including insufficient diagnoses and high therapy burdens [95], which lead to a high expenditure on healthcare. Although there hasn't been a thorough investigation of this interaction phenomenon, patient associations, healthcare websites, and personal blogs have been active on social media to increase public awareness and provide information about this difficult to diagnose rheumatic disease [84]. Patients usually use these channels to get emotional support and peer health support [96, 97], and to find treatment options or advice on healthcare decisions [98].

Several publications have previously explored the role played by people's perspective and community interactions in providing worthwhile medical decision-making information using social media analysis [99]. In different contexts and with different purposes, social media analysis has previously been used to collect information on behavioural patterns [84]. The content analysis of cancer medication uses and side effects [100] demonstrated how the internet could be a valuable tool for individuals to describe side effects and how health professionals could support medication adherence by monitoring social media discussions. Another example is found in tweets about diabetes and diets [101], which highlight how users who act as diabetes advocates can influence others' attitudes and behaviour through the propagation of information [102]. There have been other studies reporting the benefits of social media interaction and strategy for higher patient satisfaction and patient engagement, providing greater value to hospitals adopting these policies [103].

Over the past decade, social media platforms have become part of people's daily lives [114] and text mining has become a popular technique to analyse users' communication on such platforms. There have been several studies that have used social media for analysing real world events and understanding trending topics, including news events, natural disasters, and user sentiments [115]. Over millions of users use Twitter every day, and it is the most popular form of social media [113]. The streaming tools make it easy to access tweets posted online [113]. Data acquisition and research can be done using these social network portals [113]. You can create a profile, select users to connect with and view connections with by using the services [113]. Tweets are the most widely used means of exchanging content among consumers [116][117][118]. A tweet is a short, simple message that is broadcast in real-time [113]. Every user is able to see the "tweets" of all their followers", i.e., Subscribing users to their profiles. Approximately 269 million people use Twitter worldwide as of August 2019, with 37% of users aged 18-29 [113].

Influenza, according to social media analysis [119], is the most prevalent disease. By applying methods such as supervised classification, social network analysis, and linear regression, the researchers found the topic of influenza in data from Twitter [120]. In social media research, researchers have studied dental problems [121] [122], cardiac arrests, cholera, mental health, and alcohol, tobacco, and drug use. In the Twitter community, real time information is available far more frequently than surveys that may take weeks or even years to provide [113]. It is common for users to share information with doctors that they do not normally share [113]. In this case, Twitter becomes a source of new information [113]. According to the World Health Organization (WHO), 12% of the world's smokers are based in India [113]. The use of tobacco causes approximately 10 million deaths every year in India [124]. India is the second largest consumer of tobacco in the world and therefore faces a high risk of tobacco-related diseases [125]. An analysis of Twitter's large text data collection is presented in this paper to identify health-related topics [113]. Research in the past focused on identifying topics based on terms that appeared frequently in documents [113]. Low frequency topics are difficult to detect with such topic modelling [120][123]. Topic models are primarily generated based on the frequency distribution of words [113]. In the context of public health, such words are less frequent than traditional topic models [113].

Health beliefs exist in social media, and they may extend the understanding of topics such as diagnosis, medicine, and claims [130]. There are reported almost 140 potential healthcare uses of Twitter [131]. The most common uses are disaster alerting and response, diabetes management, drug safety alerts from the food and drug administration, biomedical devices data capture and reporting, shift bidding for nurses and other healthcare professionals, diagnostic brainstorming, rare diseases tracking and resource connection, smoking cessation assistance, infant care tips to new parents and post discharge patient consultations and follow-up care [131]. One example of how social networks portray medical information is the human papillomavirus (HPV) vaccine, even though its safety and effectiveness are proven. Some countries, including the United States, report low effectiveness in social networks [130]. The public trust in this topic is negatively affected by negative opinions and information caused by news, celebrities, or trendsetters [5]. There is a growing interest in social network analysis and the development of models due to the impact of social networks on healthcare [130]. In this study, [132] evaluates tweets related to health by applying machine learning techniques to regular expressions. They collect and analyse data based on regular expressions in Spain and Portugal and then narrow down the categories to four: pregnancy, depression, flu, and eating disorder [132]. In [132], KNN and SVM were the machine learning techniques used.

A model is proposed [132] based on the LDA model, and they set it up to create 250 topics, selecting "Tobacco" as a topic for validation [133]. Two other studies, one conducted in the UK and the other in the US, have also found a correlation between the sentiment analysis on twitter and quality of healthcare [134,135]. A method for automating the topic modelling of Twitter tweets is presented in this paper [130]. There will be no seeding of this system, and it will be improved based on the positive and negative feedback received [130]. By using Latent Dirichlet Allocation (LDA) as an unsupervised model, the system collects tweets, labels them, and identifies patterns [130]. Twitter tweets about public beliefs about healthcare will be processed using this method [130]. In this study, we combine CNNs with Word2Vect models [130]. In a first iteration of training [130], the Word2Vect model was trained using 7,821 medical abstracts. Training results improved the terminology related to healthcare, improved the methods to detect related tweets, and improved the general topic modelling for new tweet detection [130].

Public health research using social media has grown exponentially. Social media is used to understand the public opinion on covid-19 disease which has a high prevalence throughout the globe. Pharmaceutical companies have been releasing covid vaccines in various parts of the world through planned programmes. The vaccines efficiency and side effects can be understood by conducting on research on social media sites like Twitter, Facebook, and Instagram. Tools like topic modelling can help us understand the opinion based on a set number of topics. These topics will help us understand the public belief on a particular disease as deadly as coronavirus.

2.2.5 Topic Modelling Using social media

Topic modelling entails the unsupervised analysis and extraction of topics from a corpus of documents using a natural language processing language [84]. According to [84], this approach fit quite well with Twitter content analysis. As topic modelling is unsupervised, this method was able to identify thematic structures (topics) within a set of tweet texts without prior data manipulation, such as text labelling or training datasets [84]. With the topic modelling tool, it's possible to discover latent themes and patterns present in any text corpus, allowing you to summarize them, visualize them, and visually represent them [104]. Latent Dirichlet Allocation (LDA) is one of the most common topic modelling approaches [105]. It is a probabilistic generative model that assumes that documents consist of (latent) topics which each contain words. In this context, classification can be viewed as an alternative to numerical features or a grouping of words [84]. Structured topic modelling (STM) [106] is one of the recent applications of the LDA framework that can yield valuable results when analysing large corpora of text. By applying document-level covariates to the dataset, STM allows the user to examine metadata defined for the tweet, such as the Tweet's author, the tweet's numerical score, and various other characteristics [84]. We estimated the effect of external and network influencer scores as covariates on topic prevalence using the stm R package [107], exploring which topics appeared more frequently in tweets, and if different topics were used differently in tweet texts.

The choice of the model that is best suited to estimating the outcome of the STM is another step in the evaluation process that needs to be addressed before the final evaluation [84]. Evaluation of initialization parameters is necessary, in addition to discarding models with low likelihood values [107]. The ground truth approach is also not applicable in this instance [84]. One of the best methods of evaluating the quality of models is by comparing the intensity of semantic coherence [108] and exclusivity [106] within each topic within the model. The semantic coherence metric is a measure of the words appearing together in a topic that are most probable to occur together [84]. As part of the exclusivity metric [109], word frequency is considered. By comparing the highest scores among the topics, these measures provide a measure of the distinctness of the topics [84]. The corresponding words and topics were associated based on their (beta) probabilities of belonging to the respective topics [84]. There was no automatic generation of topic labels [84]. The label selection was the moment when researchers analysed their results after setting parameters and checked whether the allocation that emerged from their model was coherent, or if there was a need for more executions [84]. During the process of interpreting and labelling topics, each word was assigned a probability for belonging to the specific topic in question [84].

In real-time, Twitter enables millions of people to share information across the globe [84]. The sharing of valuable information and practices across online social communities allows policymakers, healthcare stakeholders, and others to influence, and be influenced by, opinions and discussions [84]. This type of possibility has proven to be extremely useful, especially for diseases of low severity and complexity, such as Lupus [110]. Through interaction among online communities, healthcare organizations can improve not only their online approaches, but also their ability to influence attitudes and behaviour [111]. The vast and fast nature of social media platforms makes it difficult to detect and depict valuable information, causing often incorrect or rumoured information to spread [112]. As far as we know, this is the first study to systematically analyse deep latent topics discussed online by online communities regarding a low-prevalence disease like Lupus [84]. The application of these methodologies to low-prevalence or rare diseases can be highly beneficial compared to other kinds of diseases, such as diabetes, HIV, or stroke, where the vast population presents more opportunities for investigation, showing inadequate care or treatment or identifying unmet needs [84]. Consequently, public health institutions should systematically examine how interactive social media features can effectively attract public attention and maintain public communication [84].

According to [113], a topic is a cluster of terms that represent a definite set of content. By using topic modelling, a collection of documents can automatically be categorized into relevant topics [113]. Topic modelling is performed using test data on tobacco and alcohol use in India and the world [113]. Tobacco use among people, primarily cigarettes and other versions on a local level, is a major health concern [113]. With Latent Dirichlet Allocation (LDA), topics of interest or those that were trending could be modelled based on the data [113]. Identifying hidden structure or themes well-represents the collection is the main challenge in topic modelling [113]. It is a method for managing, annotating, and organizing large collections of text or documents that uses a simple algorithm [113].

Latent Dirichlet allocation (LDA) is the method most used for topic modelling [116, 126]. In this case, topic modelling is performed on an extracted tweet dataset using LDA [113]. In LDA, one of the most widely used topic modelling algorithms, topics are learned by considering words that fall at the same place in documents [127]. Latent topics are distributed multinomially over the single topic. With the help of the model, the observed documents and words in a document are converted into per-document distributions of hidden topics [113]. Topic distributions are represented by P (document, topic), and word distributions by P (word, topic) [113]. The LDA model is Bayesian and based on Dirichlet distributions with hyperparameters α and β [113]. Taking the parameters into account, each word is independent [113]. They are considered unsupervised models since they self-organize words into clusters based on topics and assign documents to these topics [113]. A variant of LDA was used in the experiments in which common and non-topical words were considered additionally [113]. With this type of modelling, the results are less noisy [113]. Based on this assumption, each word originates from the background distribution via LDA with probability 'p'; for probability 1-p, a word comes from the standard LDA model [128,129].

2.2.6 Research on Covid-19 using Twitter

Concerns related to COVID-19 vaccines gained public attention as the pandemic spread globally [136]. The development of vaccines has been undertaken by a number of research teams in major pharmaceutical companies around the world [137,138]. Even though vaccinations have historically been regarded as an essential part of preventing communicable diseases [139,140,141], they have also been met with public scepticism, hesitancy, and even opposition [137,138]. COVID-19 vaccine is estimated to be necessary to provide herd immunity in a world affected by a COVID-19 pandemic [142,143], depending on the country and the infection rate. An opinion survey published in September 2020 about COVID-19 vaccine intentions suggests that 21% of U.S. adults are highly likely to be vaccinated versus 24% who are extremely unlikely to be vaccinated [144]. It is widely acknowledged that poor health literacy is one factor that determines vaccine acceptance [145]. To better understand public perceptions, concerns, and sentiments about the COVID-19 vaccine, it is imperative to understand how discussions on social media have been conducted through social media [136].

There has been extensive analysis of social media data related to health topics and emerging public health crises [146–150], but a limited comparison of big data on the conversation about COVID-19 vaccines has taken place [151,152]. The latest published study of discussions on social media related to the COVID-19 vaccine ended in November 2020, as far as we are aware [152]. The Centres for Disease Control and Prevention (CDC) confirmed that more COVID-19 variant cases occurred in North America since the vaccine was launched, vaccine deployment, and an increase in vaccines showing high efficacy since [136]. Social media discussion about vaccines has been found to reflect changes in reality according to previous research [151,153]. For a complete understanding of the public debate on COVID-19 vaccines during the pandemic, research involving recent social media data is needed [136]. Also, observing the content of social media discussions about COVID-19 vaccines may shed light on users' attitudes towards vaccines, including whether they will accept or reject the vaccine. COVID-19 vaccines had been studied previously, but their research did not address these topics [152, 154-157]. The aim of this study is to identify the topics, overarching themes, and sentiments around COVID-19 vaccines and vaccination on Twitter since the World Health Organization declared it a global pandemic on March 11, 2020, until January 31, 2021 [136].

1,499,421 unique tweets were collected in the final cleaned data set from 583,499 unique users [136]. To further improve the analysis, we removed non-analytical terms from the tweets (such as "the", "very", and "and") [136]. The stop words list was created by adding the 13 keywords associated with COVID-19 and the seven keywords associated with a vaccine to the English stop words list provided by the package tidy text, version 0.2.6. Keeping these keywords in the tweets would have no benefit to our understanding of the main content [136]. Finally, we used R packages text stem and lemma to stem and lemmatize the words (for example, we changed vaccinating, vaccinates, and vaccinated into vaccinating) [136]. We summarized our pre-processing procedures in Figure 2.2 [136].

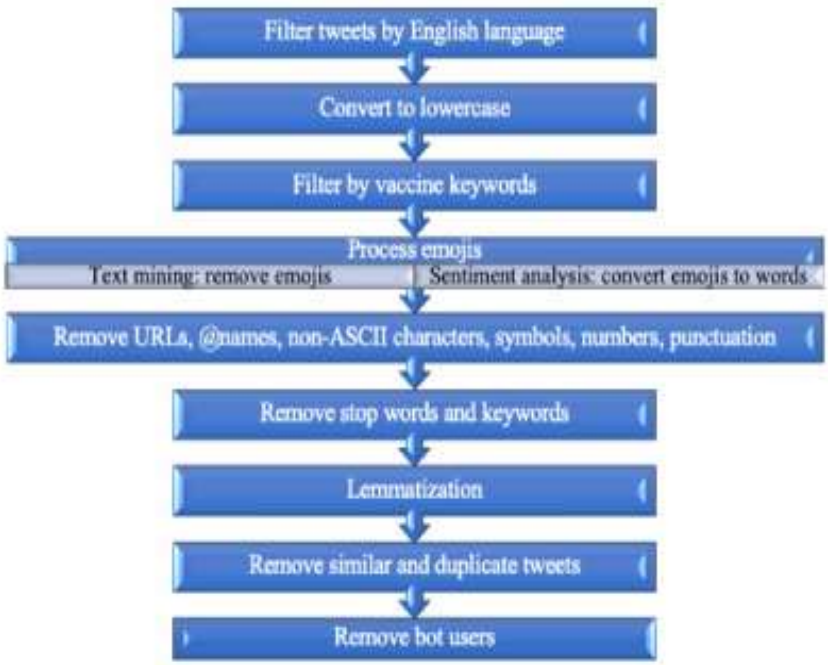

Figure 2.2. Data pre-processing procedure. ASCII: American Standard Code for Information Interchange [136].

LDA was used to model common topics based on a large collection of tweets [136]. With the R textmineR package, version 3.0.4, we performed the LDA algorithm on the data [136]. It is necessary to manually enter the expected number of topics for the LDA algorithm. In this study, the topic number was varied from 2 to 40 when running the LDA algorithm [136]. With textminR, the coherence score was calculated for every topic number [136]. Two considerations led us to select 16 topics for our final topic model: first, "16" was the topic with the highest coherence score; second, "16" struck a balance between topics that were too narrow and topics that were too broad [136]. Based on these 16 topics, each tweet was assigned a probability by the LDA [136]. Tweets were grouped according to their most prevalent topics, and the topics with the highest probability were assigned to each tweet [136].

A random sample of 100 tweets was taken from each topic, and two authors independently examined each sampled tweet, then a group discussion decided which tweets were representative for each topic [136]. A further 100 tweets would be sampled and reviewed by one of the authors if the first 100 did not reveal any notable topics; the authors continued this process until they perceived a noticeable topic and reached a consensus [136]. They categorized the topics further into five overarching themes based on their discussions [136]. In more detail, two of the authors divided the content into themes by independently determining which topics made the most sense to them and by discussing conflicting ideas [136]. During the discussion of the agreement and disagreement between the two authors, a third author contributed additional comments [136]. All three authors contributed to the final decision on grouping [136]. During the two-author discussion, it was unclear whether "vaccination drive in India" belonged to the "vaccine administration" theme or to the "vaccines as a global issue" theme [136]. As we reread tweets and discussed with each other, we eventually concluded that the topic fits within the vaccines as a global issue category [136]. A total of eight terms per topic were generated [136]. The frequency polygons (see Figure 2.3) were generated by ggplot2, version 3.3.2, for March 11, 2020, through January 31, 2021, by using the geo_freqpoly function [136].

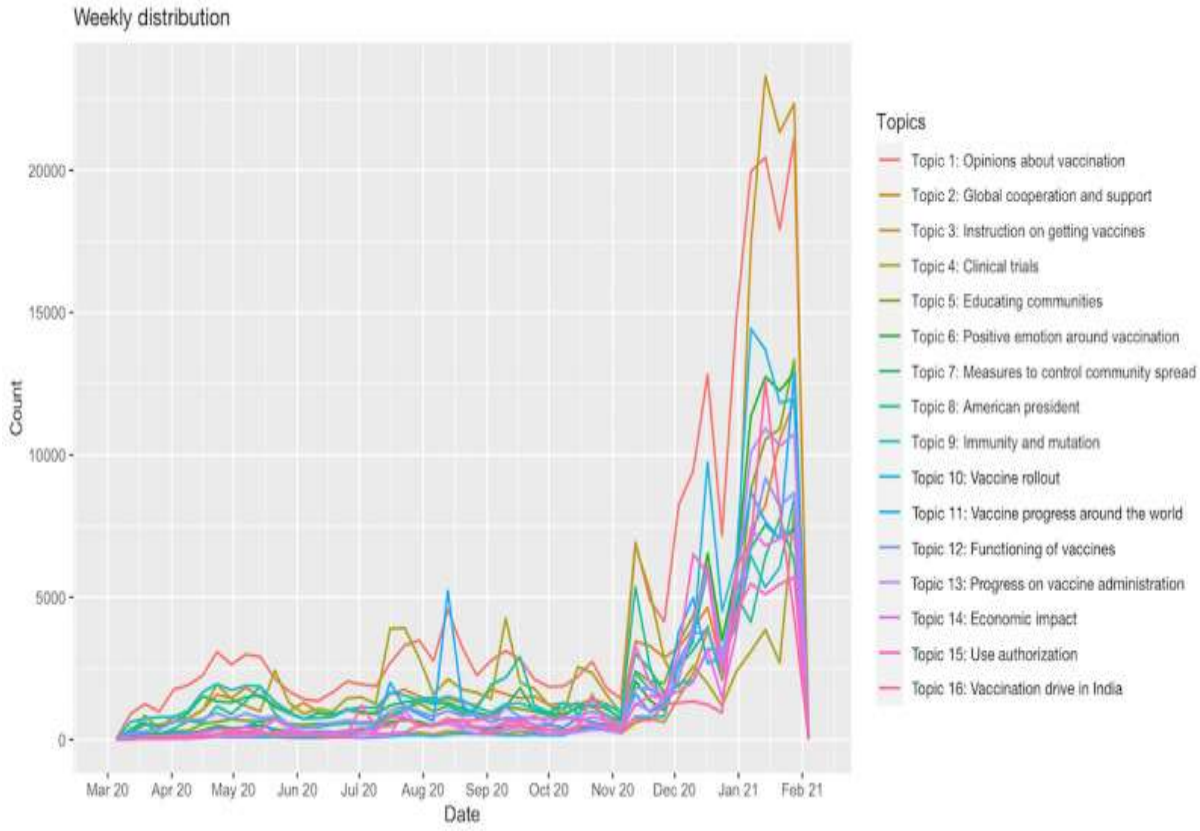

Figure 2.3. Weekly frequency of each topic on Twitter from March 11, 2020, to January 31, 2021 [136].

An opinion can be classified as positive, negative, or neutral using sentiment analysis; it can be assigned a score based on the expressed opinion. Besides simple polarity, a system of emotional analysis can result in a score for each emotion, including anger, fear, expectation, trust, surprise, sadness, joy, and disgust (as with the Plutchik wheel of emotions) [158]. As one of the most popular and efficient R packages used for sentiment/emotion analysis (Jockers, 2017) [159], syuzhet makes sentiment analysis part of an easy process. Turney and Muhammad developed the National Research Council of Canada Emotion Lexicon in 2010 [160]. In terms of comprehensiveness, it is the best choice [161]. According to Figure 2.4, the weekly average polarity (sentiment) scores between March 11, 2020, and January 31, 2021, have a slope and intercept of 0.003764 and 0.1653927 respectively with a P-value of <.001. The weekly percentage of emotions is shown in Figure 2.5.

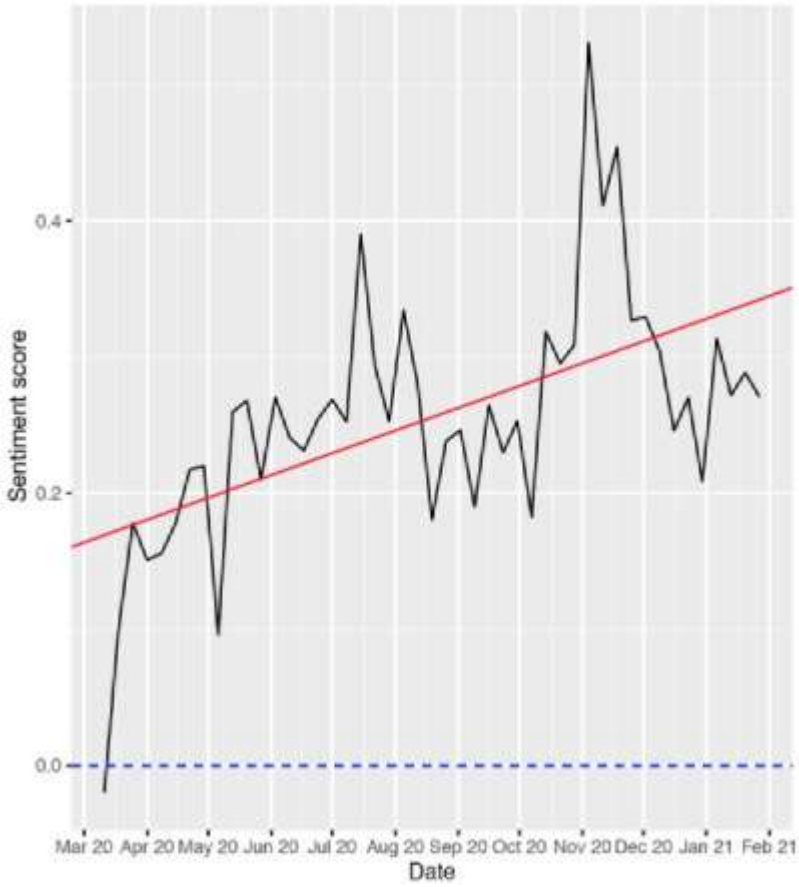

Figure 2.4. Weekly average polarity (sentiment) scores from March 11, 2020, to January 31, 2021.

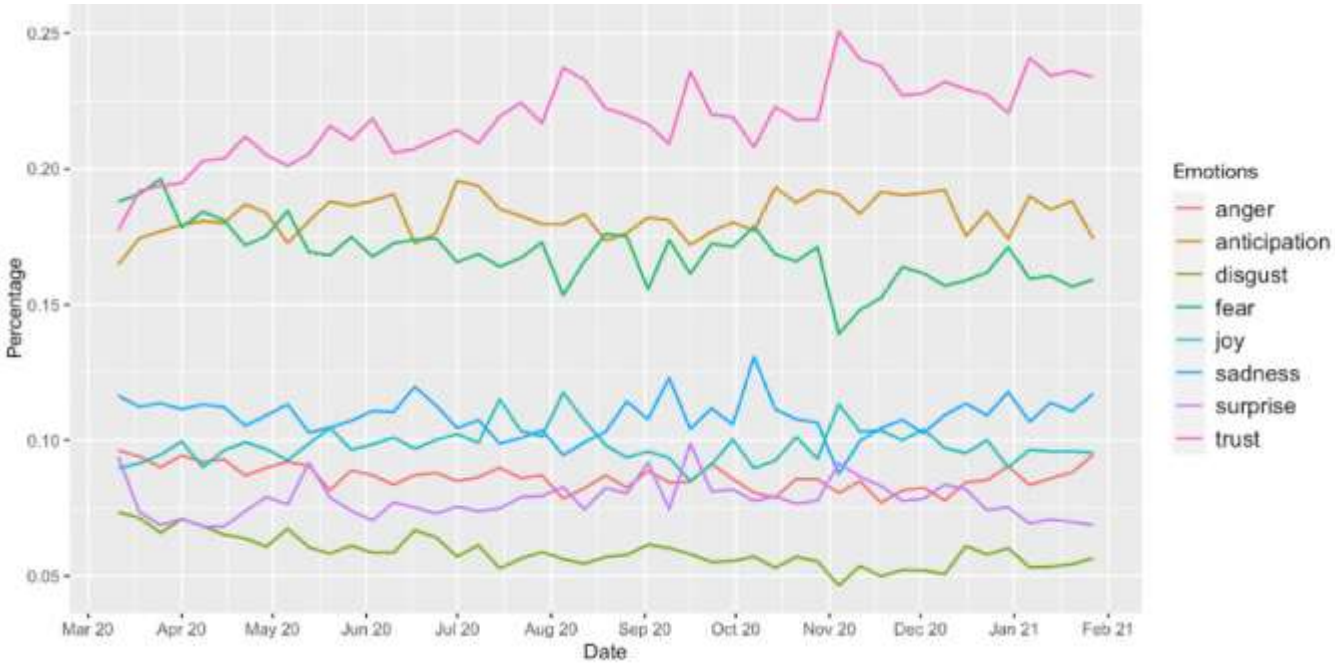

Figure 2.5. Weekly percentages of emotions from March 11, 2020, to January 31, 2021.

On average, there were 22,202 tweets per day in January 2021(Figure 2.6) a consistent increase over the previous month [136]. There were usually about 5000 tweets per day before November 9, 2020, with one exception on August 11, 2020 (n=7486), when Russia approved the first COVID-19 vaccine ever [162].

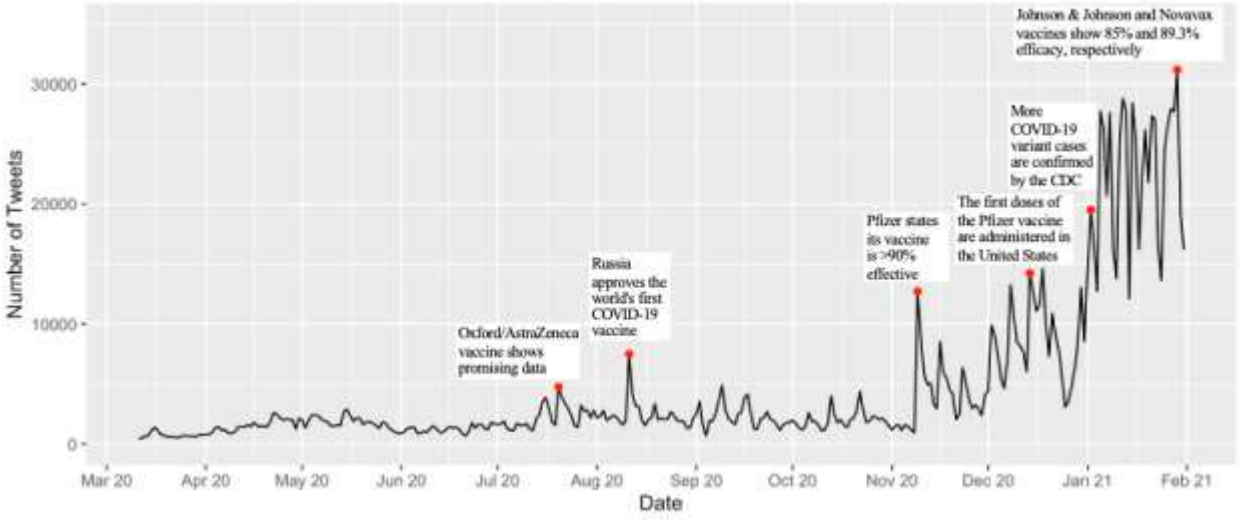

Figure 2.6. Daily numbers of COVID-19–related tweets from March 11, 2020, to January 31, 2021.

Over time, sentiment toward vaccination against COVID-19 has become more positive [136]. At the beginning of November 2020, the sentiment score reached a maximum, which coincided with the report of the high efficacy of the Pfizer vaccine [136]. Throughout the discussion, trust played an increasingly important role; it reached its peak from early November onward following Pfizer's announcement regarding the efficacy of the vaccine [136]. Overall, more people expressed trust in the discussion of the COVID-19 vaccine based on the increase in tweets. Twitter's change in fear/trust tweet percentage mirrors the change in trust tweet percentage [136]. In general, people's fear decreased as the pandemic vaccine was developed, as shown by the decrease in the overall percentage [136]. As a result of the declaration of the global pandemic in March 2020, and Pfizer's announcement shortly afterward, both of those points rose by mid-March [136]. The level of fear has decreased as vaccine studies and testing move closer to a promising outcome [136]. Over time, the percentages of other emotions in tweets remained stable [136]. A strong emotion of trust is reflected in the fact that vaccination is the only alternative [136]. Due to the prevalence of COVID-19, the speed of its spread, the disruption of normal life, and the lack of other options that have proven to be effective, vaccination has increasingly been seen as the only feasible approach to ending the pandemic [136].

A COVID-19 outbreak on a global scale necessitates a global discussion regarding the vaccine [136]. Pandemic shows how interconnected the world is now, and vaccination has become a global issue. If a country is unable to reach a certain level of vaccination for its population, there is a high risk of contagion and virus mutation [136]. Thus, the development of vaccines and the impact of the pandemic are critically important [136].

Chapter 3

Data Collection and Pre-processing

3.1 Introduction

To better understand the development of the system, this chapter provides technical background information. The next step will be to discuss data warehousing techniques. We will begin by discussing how Twitter APIs collect data. This database stores information about tweets from Twitter. After briefly discussing the pre-processing pipeline, we will discuss how it will give us clean text data.

3.2 Data Collection and Processing

3.2.1 Introduction

Twitter API was used to collect the data for this project. Using data warehousing techniques, collected data is stored and is discussed in detail in the following topics.

3.2.2 Twitter API

The API allows users to access Twitter data and analyse it, in addition to participating in the conversation. The API provides access to a variety of resources, including tweets, users, direct messages, lists, trends, media, and places [163]. A developer account must be obtained before using the Twitter API. Academic research candidates should apply to the Academic Research track. Support level, access level, and pricing are tailored to both options [164]. The academic research track gives academic researchers access to enhanced functionality, including the ability to search the full archive, a greater tweet limit, and filtered searches [164]. Upon approval and creation of the Project and App, credentials will be generated for this account. Twitter's REST API can be accessed with the tweepy Python library. Since twitter's REST API is accessible using a REST API, we can use it. A wrapper class provides easy access to Twitter REST APIs [165]. To access it, our app needs to be authorized by Twitter. Python is used to retrieve data from twitter using these credentials.

3.2.3 Storing Twitter Data

An excel workbook is used to store the data retrieved from the tweepy API. Tweets are returned by all Twitter APIs using JavaScript Object Notation (JSON) [166]. Named attributes and associated values are used in JSON as key-value pairs. An object is described by its attributes, and their states. A Tweet includes a message, an author, a timestamp, a unique ID, and sometimes location shared by the user. There are several followers on all Twitter

accounts, and each user has an account bio. The entity objects we create for each Tweet are arrays of the tweet's common contents, such as hashtags, mentions, media, and links. The columns are created in a pandas data frame and stored in a .xlsx file.

3.2.4 Data Extraction through tweepy API

Twitter data is collected using the tweepy package. This package uses the cursor API from tweepy. Parameters are passed as arguments. Arguments such as keywords and hashtags are used to retrieve the tweet. The number of tweets returned in a single run. The language of tweets to return. The since parameter is used to filter the dates from when the tweets need to be extract. The short or extended version of the tweet will be returned. As a result of the rate limit, the data collection ends, and the code is put to sleep. After the 15-minute period has ended, it starts to execute. When the code has gone into sleep mode, the user receives a wait on rate limit notification. The Twitter data can be extracted using the following code.

```
tweets = tweepy.Cursor(
    api.search,
    q=words + " -filter:retweets",
    count=3000,
    lang="en",
    since= '2021-07-16',
    tweet_mode="extended",
    wait_on_rate_limit=True,
    wait_on_rate_limit_notify=True,
).items(numtweet)
```

Figure 3.1. Tweepy API

JSON is returned as a format for the extracted tweets. Columns relevant to this JSON are extracted by iterating over it. This can be done using the following code.

```
for tweet in list_tweets:
    created_at = tweet.created_at
    username = tweet.user.screen_name
    description = tweet.user.description
    tweetid = tweet.user.id_str
    try:
        mentions = []
        for value in tweet.entities["user_mentions"]:
            mentions.append(value["screen_name"])
    except:
        print("Not working")
    location = tweet.user.location
    following = tweet.user.friends_count
    followers = tweet.user.followers_count
    totaltweets = tweet.user.statuses_count
    retweetcount = tweet.retweet_count
    hashtags = tweet.entities["hashtags"]
```

Figure 3.2 Code to extract specific columns from json

3.2.5 Storing Twitter Data

The data extracted through the API is stored in a pandas data frame in the memory. The pandas data frame was created as follows.

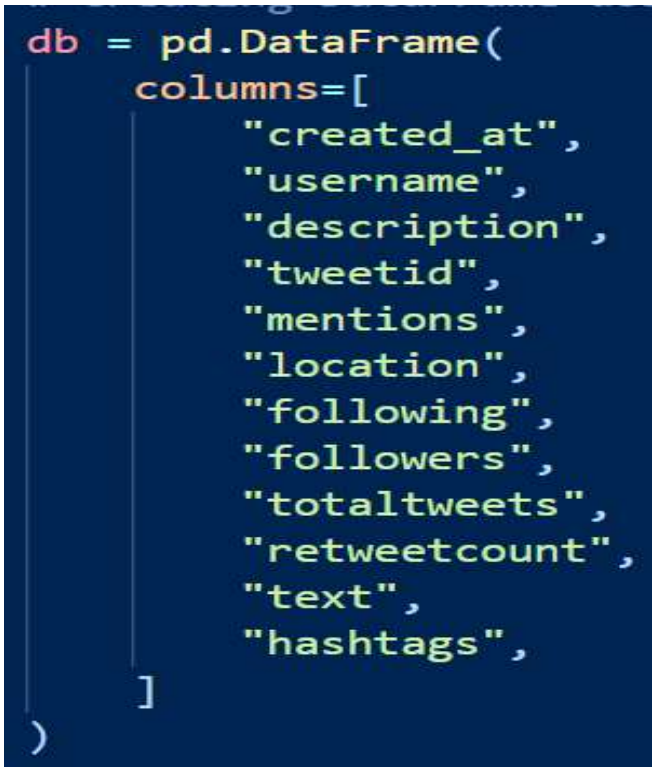

Figure 3.3 The columns of the pandas data frame

Afterwards, the data frame is saved as a .xlsx file. There is a naming convention for the file called "twitter-data-analysis-iteration-number.xlsx". The iteration is the number of iterations in the loop. Whenever an excel sheet is stored, the unique number is recorded for each run.

3.2.6 Data Dictionary

The columns extracted from twitter is described as follows.

- Created-at – The timestamp when the tweet was created.
  Example: - 2021-07-14 11:43:19
- Username – The username of the person on Twitter.
  Example: - AllisonJanel, kevojms, KerynCurtis
- Description – The description of the twitter account.
  Example: - Lover of Truth, seeker of wisdom
- Tweetid – A unique number stating the tweet id.
  Example: - 1350632570183008257
- Mentions – A list of all usernames mentioned in the tweet.
  Example: - [‘DeborahSnow', 'johncollee', 'nickzwar']
- Location – The location of the tweet.
  Example: - Dublin

- Following – The number of accounts the twitter handle is following.
  Example: - 1346
- Followers – The number of accounts who follows the twitter handle.
  Example: - 1517
- TotalTweets – The total number of tweets from the twitter handle.
  Example: - 4170
- Text– The full text on the tweet.
  Example: - “Here in CA the government is literally offering”
- Hashtags – The hashtags mentioned in the tweet.
  Example: - ['CovidVaccine', 'EUDigitalCovidCertificate']

The aggregated and final data frame.

created_at	username	description	tweetid	mentions	location	following	followers	totaltweets	text	hashtags
2021-07-14 11:43:19	jesus_reigns247	Lover of Truth, seeker of wisdom, Christ's bon...	1350632570183008257	[]	NaN	172	31	424	In my cold dead arms#COVID19 #CovidVaccine wi...	['COVID19', 'CovidVaccine', 'tyrants', 'vaccin...
2021-07-14 03:17:43	AllisonJanel	Conservative in California. Mama 🦁 to 2 babies...	358689571	[]	LA County / California	1346	791	4170	Here in CA the government is literally offerin...	['CovidVaccine']
2021-07-13 19:18:35	kevojms	Values humanity with care especially the silen...	71753501	[]	Dublin	505	763	8271	Nice to see my friend John, pushing boxes at t...	['CovidVaccine', 'EUDigitalCovidCertificate']
2021-07-12 20:14:43	TribuneCeylon	Welcome to https://t.co/tvLDEMAYIz on Facebook...	1388694554540142595	[]	NaN	18	9	1352	Online portal introduced for Covid vaccine app...	['lka', 'news', 'srilanka', 'ceylontribune', '...
2021-07-13 21:14:35	KerynCurtis	Writer/editor/communications/policy/campaigner...	27598839	['DeborahSnow', 'johncollee', 'nickzwar']	Australia	2139	1517	2461	It's #BastilleDay rv & I celebrate by shar...	['BastilleDay', 'COVIDVaccination', 'GetVaccin...

Figure 3.4 The final data frame

3.2.7 Data Pre-processing

This section will discuss the data pre-processing steps associated with the project.

3.2.7.1 Text Pre-processing

Some tweets come with emoji included in between the texts. These emojis can not be handled as it as. These emojis are converted into strings. To illustrate this further, the emoji 😂 is converted to face with tears of joy and 🤣 is converted to rolling on the floor laughing. This is done using the emot package in python.

Hashtags, mentions, any hyperlinks and emojis are removed. The 50 most common words are removed from the tweets.

3.2.7.2 Tokenization

Tokenization is a vital step in text analysis that is as basic as it is important. Using tokenization, we manipulate streams of text by dividing them into smaller units, typically words or phrases. The tweets will be tokenized with the NLTK Python library. By tokenizing the text, we prepare it for the next step, which involves removing stop-words like 'the', ‘or’, 'to’, ‘and' etc.

3.2.7.3 Removing Stop Words

Pre-processing steps include stop-word removal, an important step. There are many stop-words in every language. They are important in the language, but when taken out of context, they rarely carry much meaning. These stop words include articles, conjunctions, some adverbs, etc. Language-specific stop-words may be provided by some libraries. A default set of stop-words is provided by the NLTK library. In terms of frequency analysis, stop words are not of any value to us. It might be considered that these examples are the most frequent terms used by the English language.

3.2.7.4 Lemmatization

Stratification of a word by gathering its inflected forms together is called lemmatization in linguistics [167]. It can be done by identifying the lemma or dictionary form of each word as a single item for analysis [167]. Generally, lemmatisation refers to the process of determining a word's meaning from its lemma through algorithms [167]. Stemming is based on identifying words according to their intended function, but lemmatization requires placing words correctly within the context of the larger sentence, such as others nearby or even the entire document [167]. Hence, it is a question of research to develop efficient lemmatisation algorithms [167].

After all the pre-processing the data frame looks as given below. The lemmas back to text is the column which we will use for further topic modelling.

	text_original	emoji	tweet	clean_tweet	tokens	tokens_back_to_text	lemmas	lemmas_back_to_text	lemma_tokens
0	In my cold dead arm\n#COVID19 #CovidVaccine wi...		In my cold dead arm\n will never be in my bod...	cold dead arm never body stop govt overreach p...	[cold, dead, arm, never, body, stop, govt, ove...	cold dead arm never body stop govt overreach p...	[cold, dead, arm, body, stop, govt, overreach,...	cold dead arm body stop govt overreach poison ...	[cold, dead, body, stop, govt, overreach, pois...
1	Here in CA the government is literally offerin...		Here in CA the government is literally offerin...	ca government literally offering free weed get...	[ca, government, literally, offering, free, we...	ca government literally offering free weed pre...	[government, literally, offer, free, weed, pre...	government literally offer free weed pretend	[government, literally, offer, free, weed, pre...
2	Nice to see my friend John, pushing boxes at t...		Nice to see my friend John, pushing boxes at t...	nice see friend john pushing boxes start hard ...	[nice, see, friend, john, pushing, boxes, star...	nice see friend john pushing boxes start hard ...	[nice, friend, john, pushing, box, start, hard...	nice friend john pushing box start hard work e...	[nice, friend, john, pushing, start, hard, wor...
3	Online portal introduced for Covid vaccine app...		Online portal introduced for Covid vaccine app...	online portal introduced covid vaccine appoint...	[online, portal, introduced, appointments, wp]	online portal introduced appointments wp	[online, portal, introduce, appointment, wp]	online portal introduce appointment wp	[online, portal, introduce, appointment]
4	It's #BastilleDay rn & I celebrate by shar...		It's & I celebrate by sharing this fabul...	amp celebrate sharing fabulous french vaccin...	[celebrate, sharing, fabulous, french, ad]	celebrate sharing fabulous french ad	[celebrate, share, fabulous, french, ad]	celebrate share fabulous french ad	[celebrate, share, fabulous, french]

Figure 3.5 The pre-processed data frame

Chapter 4

Exploratory Data Analysis

4.1 Introduction

An exploratory analysis of Twitter data has been shown in this chapter. Our first step will be to look at the total number of tweets sorted by creation date. Following that is an hourly breakdown of tweets. We examine the number of followers and following accounts. By analysing their activity, we examine the twitter accounts. Afterward, we explore hashtags and mentions that are most popular. Our next step is to examine each tweet's length. We conclude our analysis by examining the location from which the tweet was posted.

4.2 Number of Tweets

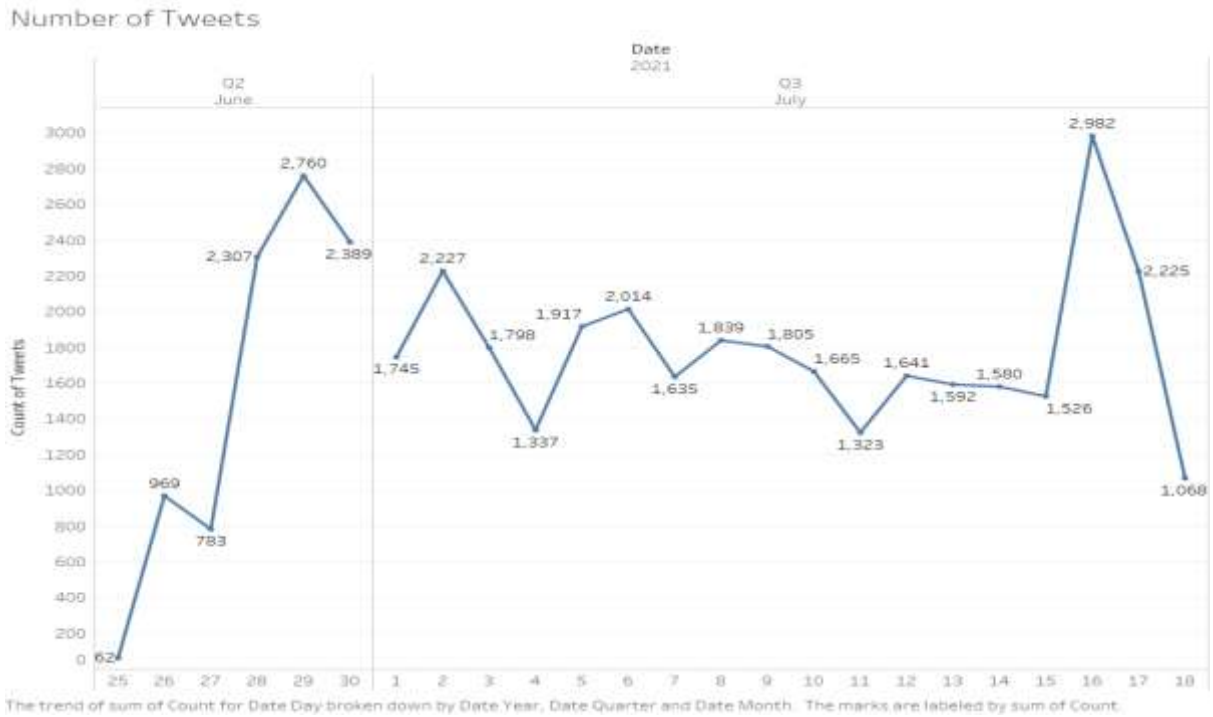

Figure 4.1 The number of tweets by date

The figure 4.1 illustrates the count of tweets collected on each day. The data was collected from the 25th of June 2021 until 18th July 2021. The x-axis is the date axis. The y-axis is the count of tweets. The labels shown on the graph are the count of tweets on that day. We see that most of the tweets were on 16th of July, while 29th June trails. The tweet count from 1st July to 15th July almost remains constant.

4.3 Count of Tweets by hour of day

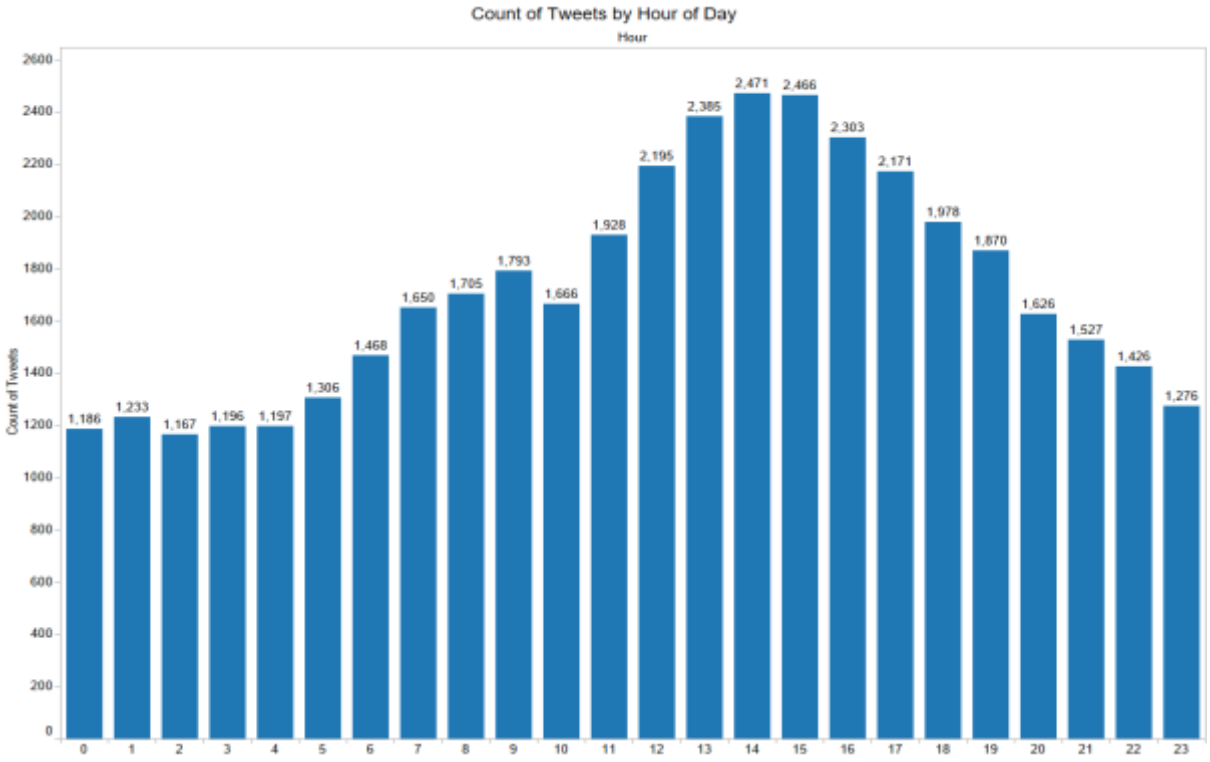

Figure 4.2. The count of tweets by hour of day

The figure 4.2 illustrates the count of tweets by hour of day at which it was posted. A day consist of 24 hours and the x-axis has the range from 0th hour until 23rd hour. The tweets are binned by the hour they belong to. The y axis is the count of tweets for the hour. We see that the busiest time was from 11am until 7pm. This is the period which has posted the maximum number of tweets on the covid vaccine topic.

4.4 Number of Followers

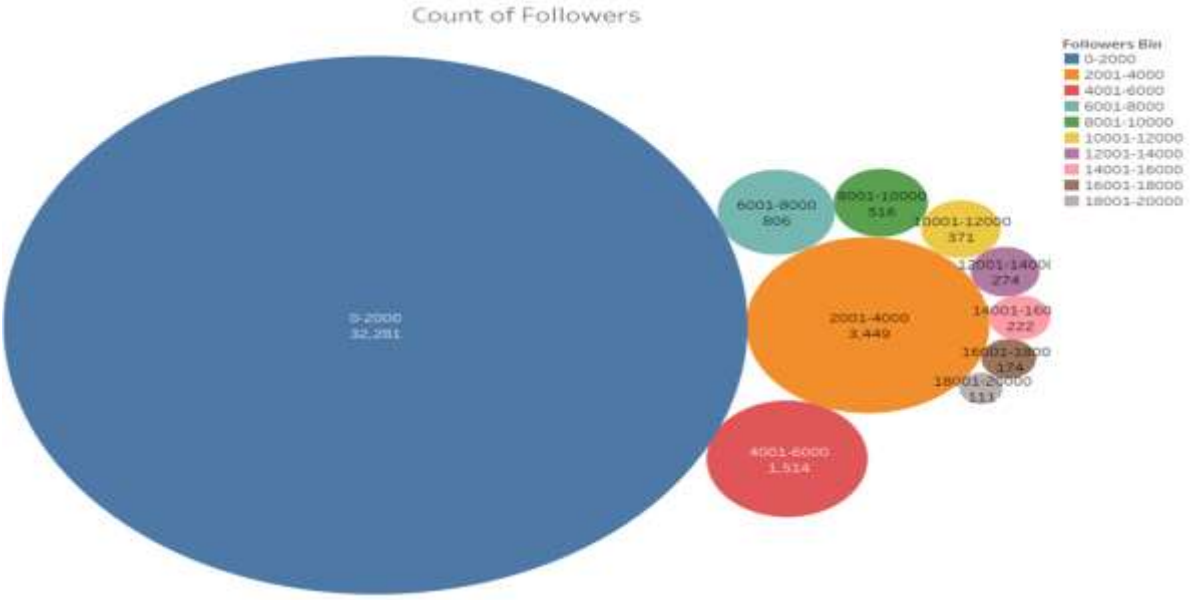

Figure 4.3. The number of followers

The figure 4.3 shows the number of followers of the twitter handles collected in the data. The followers are binned into bins of range 2000. We infer from the graphs that the most of our population consists of 0-2000 followers. While the second most popular category is the 2001-4000 followers.

4.5 Number of Following Accounts

Figure 4.4 The number of accounts following

The figure 4.4 shows the number of accounts the twitter account follow. The number of accounts is binned in the range of 500. The x-axis is the bins of the number of accounts being followed by the twitter handles. The y-axis is the count of the handles being followed in the range. Most of the samples fall into the 0-500 following bin. It is followed by the 501-1000 bin. There is another interesting point in the graph. It is the increase in the following account falling in the bin 4501-5000. This is due to the accounts which follow people extensively.

4.6 Twitter Accounts by Activity

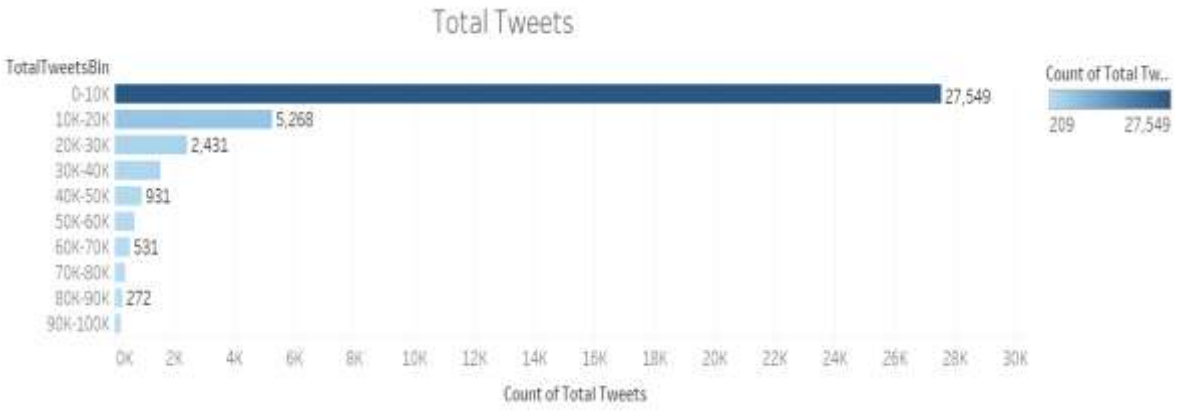

Figure 4.5 The total number of tweets

The figure 4.5 shows the activity of the twitter handles. This is visualised by the total number of tweets by the accounts. The tweets are binned in the range of 0 to 1 lakh with 10000 distances between the bins. We infer most accounts fall in the range of 0 to 10000 tweets. 27549 accounts fall into this category. The 10000 to 20000 tweets are second highest with 5268 accounts falling into the category. The high-density blue colour helps us understand the densest category.

4.7 Most Common Hashtags

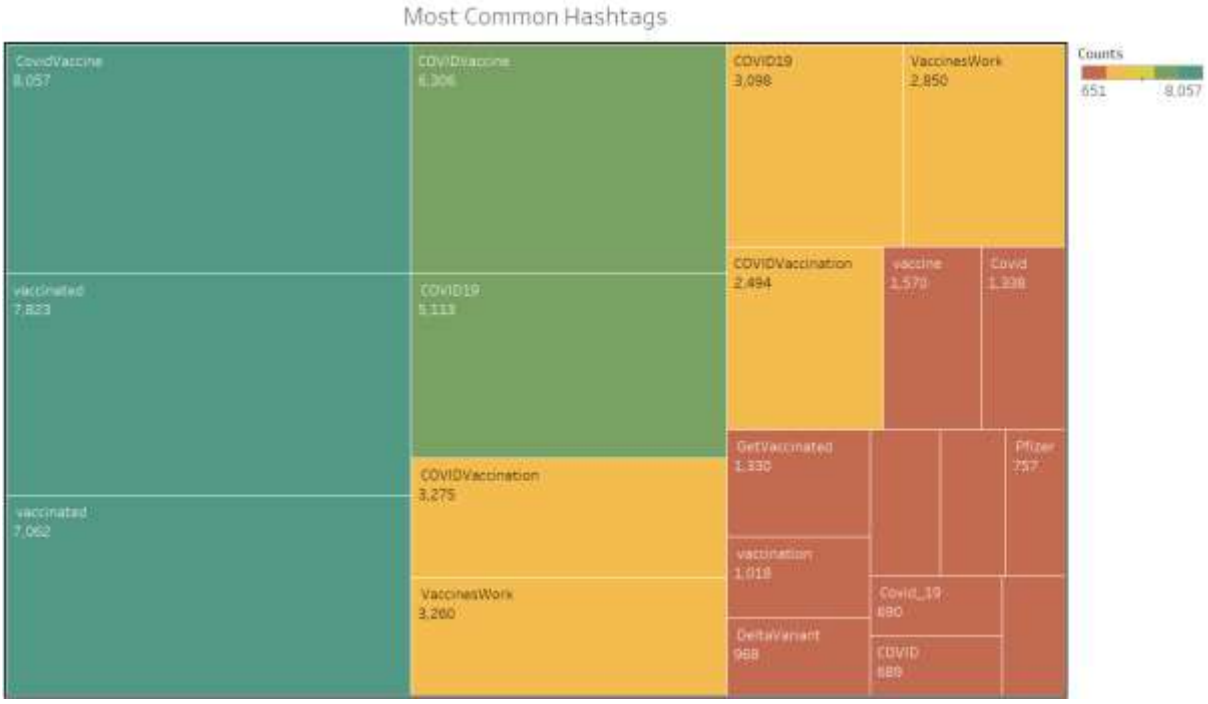

Figure 4.6 The most common hashtags

The figure 4.6 shows the most common hashtags used in the tweets collected in the database. The size of the block determines the count of the hashtags. The colour of the block determines the intensity of the number of tweets using the hashtags. Bluish green is the highest, then comes the green, yellow, and red blocks. We infer the most used hashtags in the collected data. These are #covidvaccine, #vaccinated, and #covidvaccine. The tweets contain these hashtags more as the data collection was based on these hashtags. The data is biased on these hashtags.

4.8 Most Common Mentions

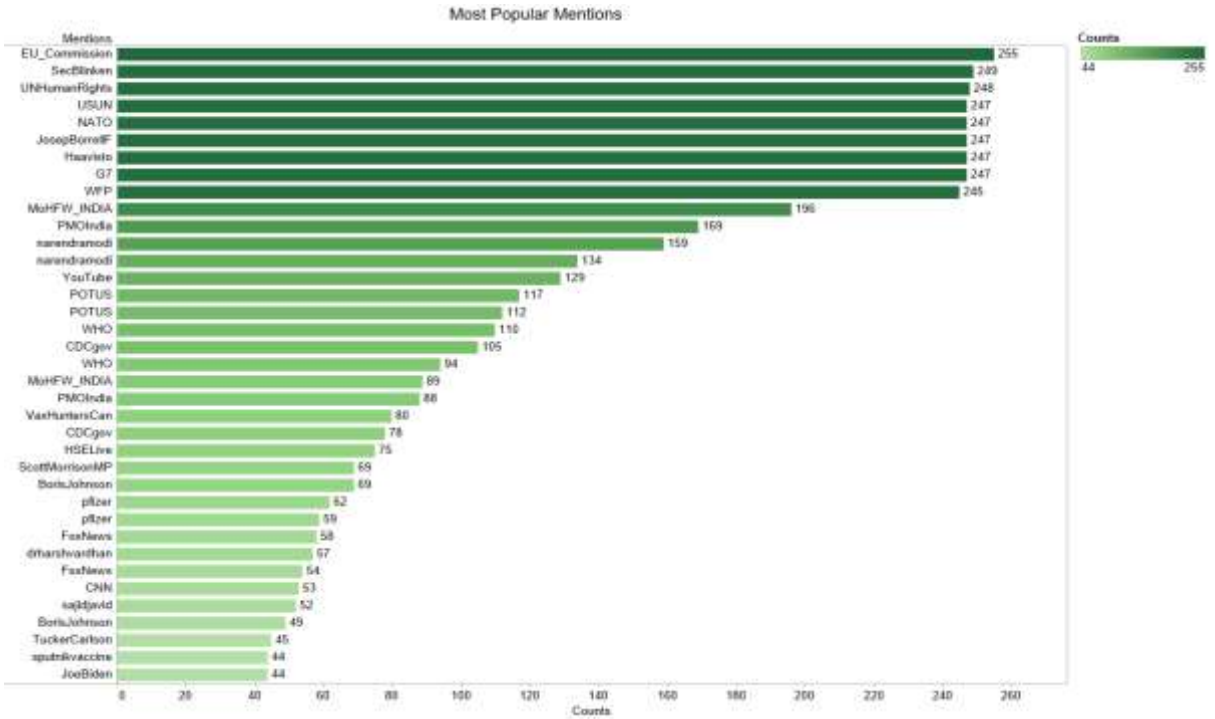

Figure 4.7. Most Common Mentions

The figure 4.7 illustrates the most common mentions found in the database. We infer that EU-commision, UNHumanRights, USUN, and NATO are the most mentioned in the covid tweets database. The counts are shaded with the colour green representing the density of each mention.

4.9 The length of Tweet

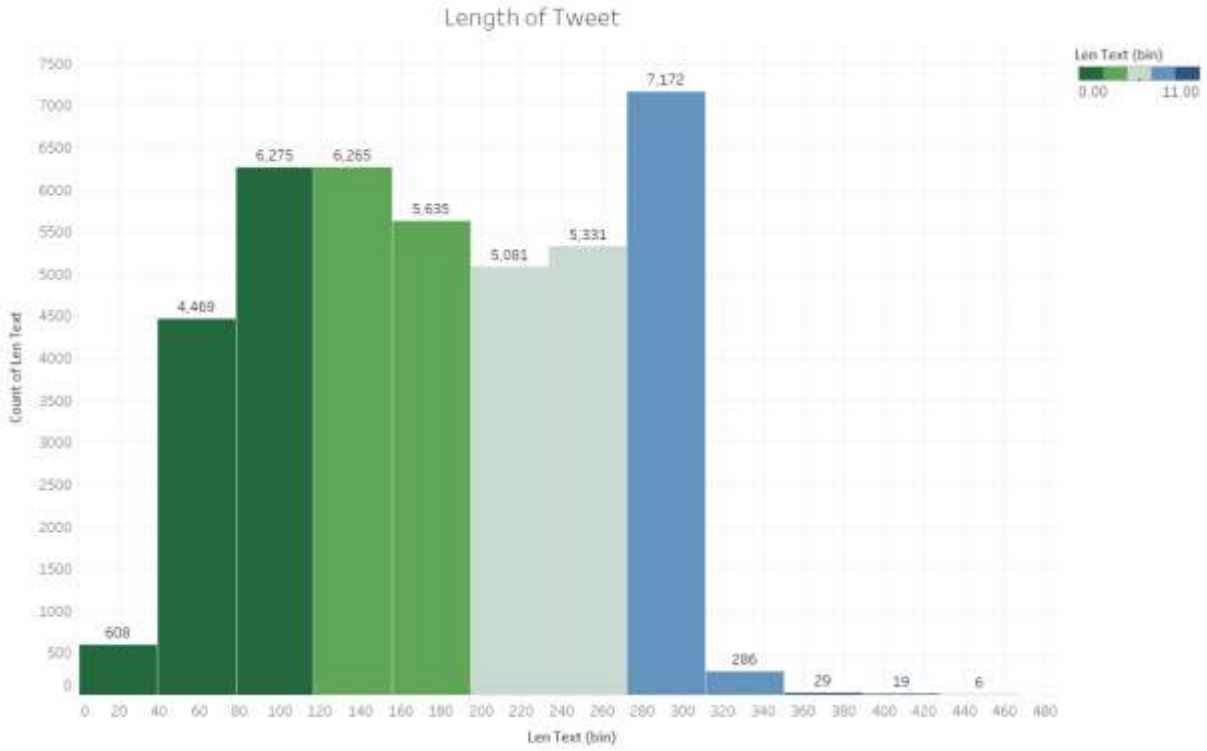

Figure 4.8. The length of the tweets

The figure 4.8 shows the length of the tweets collected in the database. We see that most of the tweets ranged anywhere between 280 and 320 characters. The second most important range of the length of tweets are from 80 to 200 characters. There are very few tweets in the range 320 characters and above.

4.10 Tweets by Geographic location

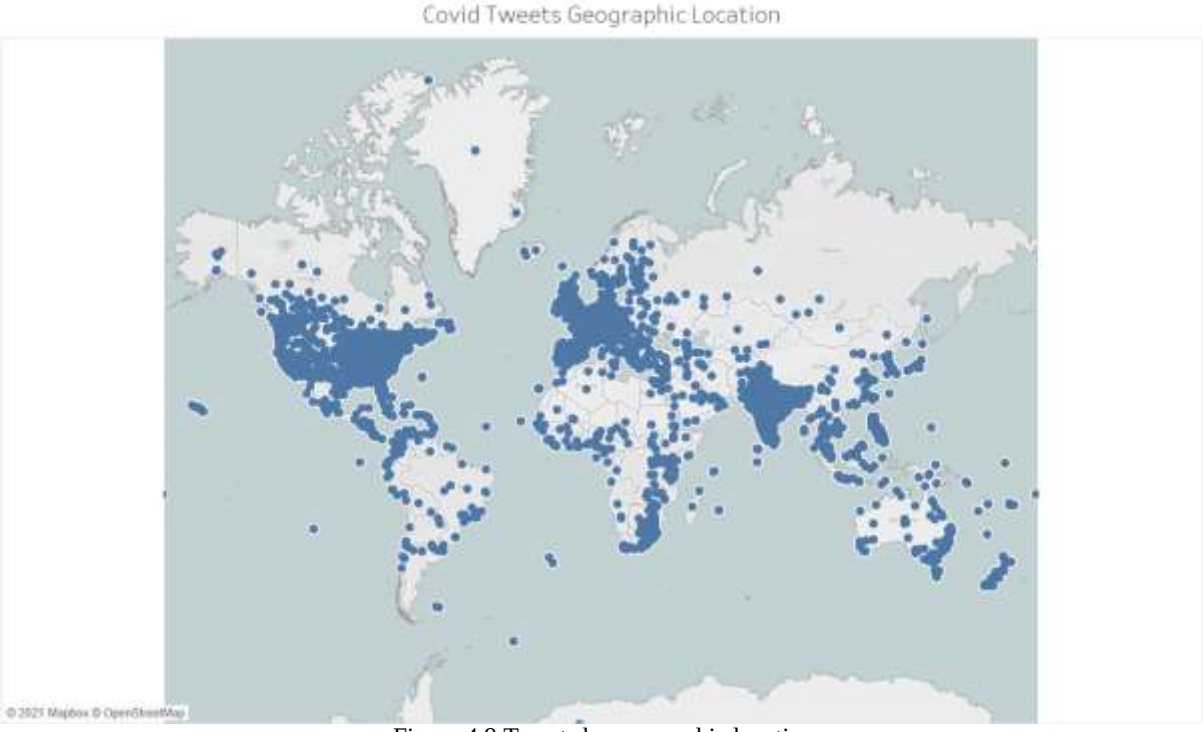

Figure 4.9 Tweets by geographic location

The figure 4.9 illustrates the locations from where the tweets have been posted from. This graph takes in the latitude and the longitude information of every tweet and plots the point on the world map. We see that we have tweets from all over the world. While still there are very few tweets from countries like Russia, China, Greenland, Northern Canada, and South America.

Chapter 5

Sentiment Analysis

5.1 Introduction

We explore how sentiment is extracted from tweets in this chapter. Depending on their content, tweets are categorized into positive tweets, neutral tweets, and negative tweets. This can be achieved by using the textblob module. Python 2 and 3 versions of Textblob use a text processing library to process text. APIs are available for a variety of NLP tasks, including phrase extraction, parts of speech tagging, and sentiment analysis.

5.2 Extracting Sentiments from Twitter

5.2.1 Sentiment Analysis: Introduction

We can generate insights into the context when conducting a sentiment analysis by examining people's moods and emotions [168]. Using a sentiment analysis, data is analysed and categorized to meet the needs of a research. With these sentiments, one can comprehend a range of events and the impacts they cause [168]. In the research literature, L Bing [169] explains that sentiment analysis or opinion mining go by many different names, e.g., "behaviour analysis, emotional analysis, affect analysis or perception analysis", however they all have similar purposes and are all related to sentiment analysis or opinion mining. We can gain valuable insight into people's values, their wants, and their top concerns by analysing these sentiments [169].

The Textblob library provides NLP functionality in Python. Detecting polarity and subjectivity in a tweet is based on this. NLP tasks are actively performed with Textblob. There are several lexical resources in NLTK, and users have access to them all easily. NLTK allows them to categorize, classify, and do much more. There is no need to create a complex toolchain for testing textual data. TextBlob is an easy solution for performing complex textual analysis. An approach that uses lexicons to understand a sentiment measures the intensity and semantic orientation of the words. A predefined dictionary is necessary to classify positive and negative Text messages are typically represented by bags of words. By taking the average of all the sentiments after assigning individual scores to each word, the final sentiment is calculated [168].

The polarity and subjectivity of a text are returned with TextBlob. In terms of sentiment, [-1,1] -1 represents negative sentiments, while 1 represents positive ones [168]. When a word is negative, the polarity is reversed. With TextBlob, you can fine-tune analysis by including semantic labels. The use of emoticons, exclamation marks, or emojis, for example [168]. Subjectivity lies between [0,1]. In text, subjectiveness refers to how much factual on information is present. It is more personal and opinionated than information because of the higher subjectivity [168]. The intensity parameter is new to TextBlob. 'Intensity' is used to calculate subjectivity in TextBlob [168]. Each word is modified by its intensity [168]. An adverb ('very good') is a modifier in English. As an example: We calculated polarity and subjectivity for the phrase "I don't like this example at all, it's boring." [168]. For this example, the polarity value is -1 and the subjectivity value is 1, which is acceptable [168]. I prefer another example for the sentence "This was a good example, but I'd like to see some other examples" [168]. In terms of subjectivity and polarity, it returns a value of 0.0, which isn't the best answer you might expect [168].

An objective statement expresses an individual's thoughts, views, or beliefs. The range of subjective values is zero to one. The number 1 is extremely subjective, while the number 0 is extremely objective. Tweets are categorized by their degree of negativity or positivity. Among the polarities, 1 is the most polar and -1 is the least polar. A positive sentence is a 1, while a negative sentence is a -1.

```
text_df['polarity'] = text_df['lemmas_back_to_text'].apply(lambda x : TextBlob(x).sentiment.polarity)
text_df['subjectivity'] = text_df['lemmas_back_to_text'].apply(lambda x : TextBlob(x).sentiment.subjectivity)
```

Figure 5.1 Code for calculating polarity and subjectivity

Column "lemmas_back_to_text" shows the pre-processed tweets. In the data frame, each tweet is subject to the textblob function. The polarity and subjectivity columns contain these values, it is illustrated in Figure 5.2. A screenshot of the code to extract polarity and subjectivity can be seen in Figure 5.1.

	lemmas_back_to_text	subjectivity	polarity
0	cold dead arm body stop govt overreach poison ...	0.700000	-0.400000
1	government literally offer free weed pretend	0.800000	0.400000
2	nice friend john pushing box start hard work e...	0.513889	0.102778
3	online portal introduce appointment wp	0.000000	0.000000
4	celebrate share fabulous french ad	0.500000	0.200000

Figure 5.2 The data frame with polarity and subjectivity columns

5.2.1 Sentiment Analysis: Trends

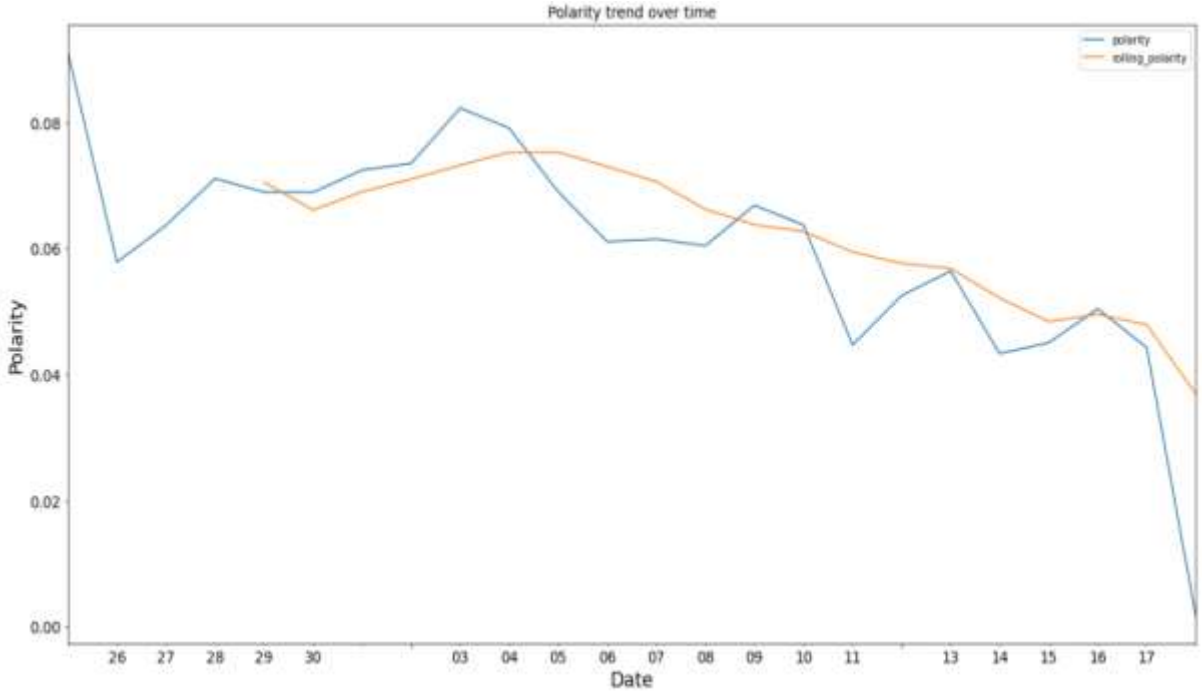

Figure 5.3 Polarity trend over time

We can see the polarity across time in figure 5.2. Our x-axis shows the data collected from 26th June 2021 to 17th July 2021. Using the textblob function, we were able to achieve a score of 'Polarity'. Graphs are depicted by a blue line indicating the average polarity for the extracted data. The orange line represents the rolling average of the past 5 days. As a result, this only appears on the fifth day of the graph. We can see where the trend is going by this rolling average. Based on the graph, we can infer that the slope is decreasing as the days pass.

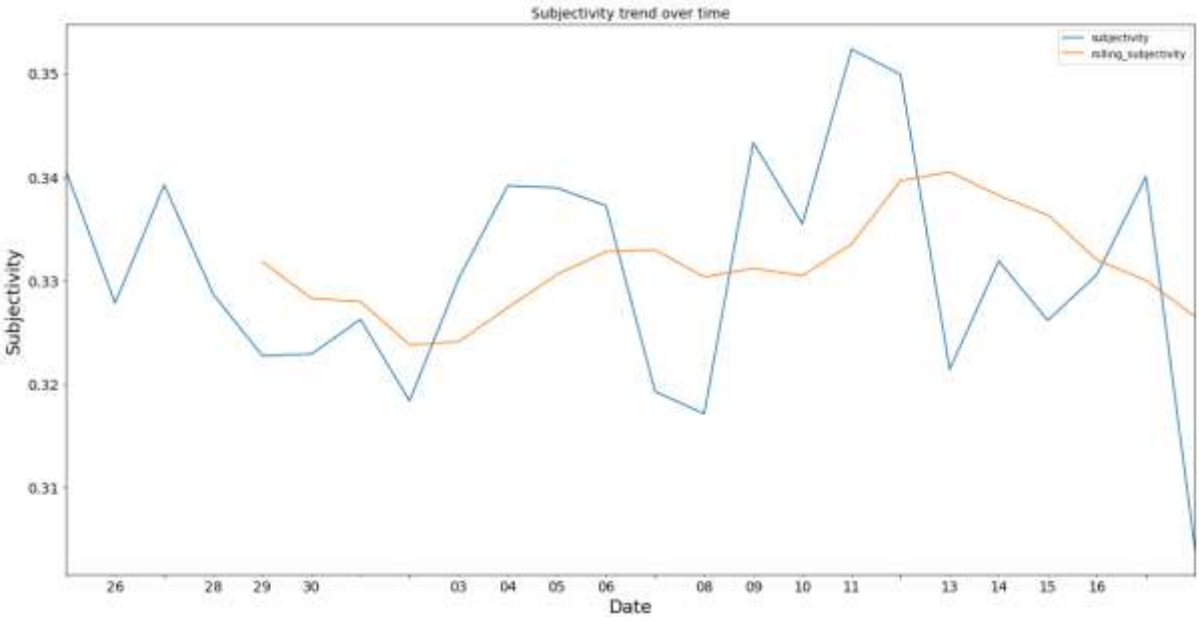

Figure 5.4. The subjectivity trend over time

A chart of the subjective trend in figure 5.4 is shown. According to this line, the day's subjectivity lies somewhere in the blue area. Using a rolling period of 5 days, the orange line represents the rolling subjectivity. From the fifth day onward, the orange line appears. We can thus understand how the graph is trending. A subjectivity value ranges between 0.32 and 0.36 approximately. Therefore, most tweets seem to be within the objective range. Some days are more subjective than others. Nine, eleven, and twelve July are these dates. For a brief period, the rolling subjectivity shows that the subjectivity increases and then decreases.

The tweets are classified into positive, negative, and neutral. This is determined by the polarity. If the polarity is less than zero it is classified as negative tweet. If the polarity is zero, it is classified as neutral tweet. If the polarity is greater than zero it is classified as positive tweet. The code used to classify is illustrated in figure 5.5.

```
def getAnalysis(score):
    if score < 0:
        return 'Negative'
    elif score == 0:
        return 'Neutral'
    else:
        return 'Positive'

text_df['Analysis'] = text_df['polarity'].apply(getAnalysis)
```

Figure 5.5. Classification of tweets

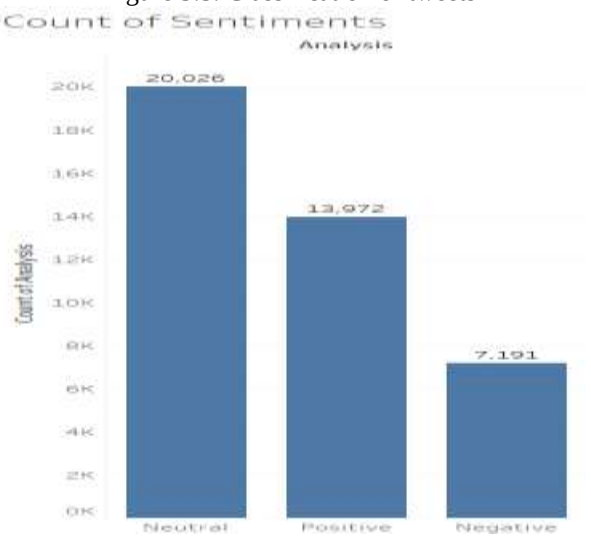

Figure 5.6 Count of Sentiments

Here's a shot of the sentiment count across the entire extracted dataset shown in figure 5.6. There were 40,000 tweets extracted from the extracted data. Twenty thousand two hundred and twenty six tweets are considered neutral, which is half of the tweets. The positive tweets number 13,972, while the negative tweets number 7,191.

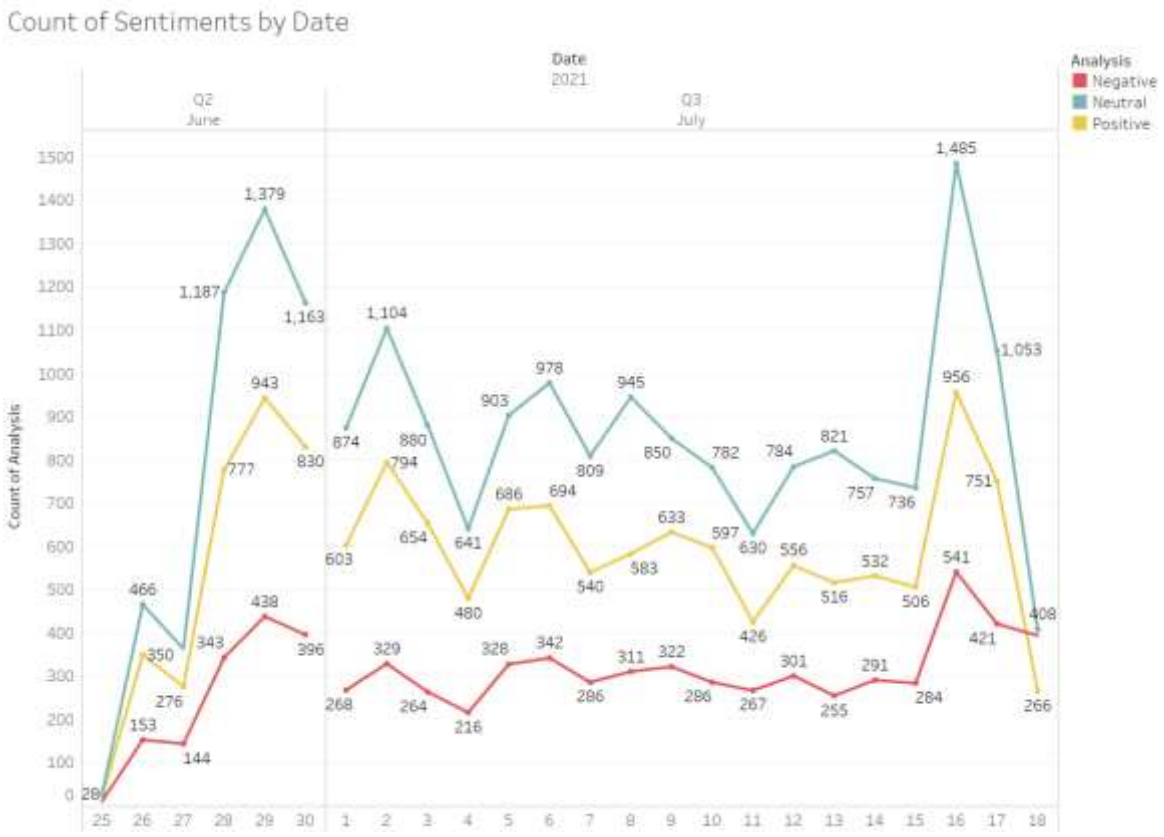

Figure 5.7. Count of sentiments by date

Figure 5.7 shows how many sentiments were expressed each day. Every day, we have seen a high number of neutral sentiments. The count of positive tweets follows, except on 18th July 2021. Each day, there are about 300 negative tweets.

Chapter 6

Topic Modelling

6.1 Introduction

We dive deep into topic modelling in this chapter. The tweets will be categorized into topics this way. Grid searching is also used to determine the ideal number of topics and the learning rate. Additionally, a visual representation of the overall term frequency of all the words related to the topic is provided. Each topic can be determined by reading the words that are associated with it.

6.2 Extracting Topics from Corpus

6.2.1 Topic modelling: Introduction

Unsupervised machine learning techniques, such as topic modelling, can scan a collection of documents, detect word and phrase patterns, and cluster them according to similar expressions to best describe the collection [171]. Using machine learning, topic modelling uses text data to identify clusters of words based on the document content. A large amount of training data isn't required because human-structured data doesn't need to be classified beforehand, so unsupervised machine learning is what it is [171]. It's easy and quick to start analysing data with topic modelling since it doesn't require training [171]. Despite this, you can't be certain you'll get accurate results, which is why many companies invest in training topic classification models before implementing them [171]. By identifying patterns and recurring words in a set of customer reviews, topic modelling could be used to identify the topics of the reviews [171]. To infer topics within unstructured data, theme modelling counts words and group them into similar word patterns. A topic model clusters feedback similar in structure and words that appear frequently based on patterns such as word frequency and distance between words [171]. These parameters enable you to deduce what each set of texts is saying very quickly [171]. It is an 'unsupervised' approach, which means that no training is required.

Latent Dirichlet Allocation (LDA) and LSA are based on the same underlying assumptions: the distributional hypothesis, (i.e., similar topics make use of similar words) and the statistical mixture hypothesis (i.e., documents talk about several topics) for which a statistical distribution can be determined [171]. The purpose of LDA is to map each document to a set of topics covering a significant number of words in that document [171]. LDA takes two inputs. One is integer mappings of each word, which is id2word. Secondly, the bag of words input. Figure 6.1,6.2 illustrates the two inputs required for topic modelling.

```
id2word = Dictionary(text_df['lemma_tokens'])
print(len(id2word))
```

```
27651
```

Figure 6.1 The integer mappings(id2word) of each word

```
corpus = [id2word.doc2bow(d) for d in text_df['lemma_tokens']]
```

Figure 6.2 The bag of words

6.2.2 Topic modelling: Grid Search

Grid search is performed using Ldamulticore. The Ldamulticore API is a system that parallelizes and speeds up model training by making use of all the CPU cores. Training with Ldamodel is slow compared to training with Ldamulticore. It iterates through a variety of topics using a grid search method. From 0.05 to 1, there are various learning rates illustrated in Figure 6.3. Three to 13 topics are iterated. A for loop is used to execute this.

```
alpha_val = [0.05, 0.1, 0.3, 0.5, 0.8, 1]
MulLda_alphas = []


for topics in range(3, 15, 2):
```

Figure 6.3 Learning rates and number of topics

The ldamulticore API is used to run the topic modelling for the tweets. This is illustrated in figure 6.4. There are many parameters used to run the grid search. The corpus is the bag of words implementation for all the tweets. The id2word is the integer mappings for each word. Random state is used to keep the results constant and not change with every run. The num_topics is used to iterate over the number of topics in each iteration. The alpha parameter is used to iterate the different learning rates for each iteration of grid search.

```
lda_model_multi_notts = gensim.models.LdaMulticore(corpus = corpus,
             id2word = id2word,
             random_state = 42,
             num_topics = topics,
             passes=10,
             chunksize=512,
             alpha=alph,
             offset=64,
             eta=None,
             iterations=100,
             per_word_topics=True,
             workers=6)
```

Figure 6.4 Ldamulticore API with parameters

In Topic Coherence, words that are high scoring within a topic are considered similar based on their semantic association [172]. These measurements are useful for identifying topics that are semantically interpretable and those that represent artifacts of statistical analysis [172]. When statements or facts are supported by each other, they are claimed to be coherent [172]. An interpretation that covers all or most of the facts can be made of a coherent set of facts [172]. After running the grid search for all learning rates and number of topics a data frame is created with the coherence values in descending order. Figure 6.5 illustrates the data frame. We infer that the best values of number of topics and learning rates are 9 and 0.05 respectively.

MulLda_Topic_Num	MulLda_alpha_val	MulLda_Coherent_score	MulLda_Perplexity_val
9	0.05	0.368453	-8.070022
11	0.05	0.349216	-8.101546
9	0.10	0.340201	-8.050964
11	0.10	0.335908	-8.076597
11	0.30	0.334519	-8.033728

Figure 6.5 Results after grid search

6.2.3 Topic modelling: Final model

The best parameters search from grid search has given us the best model parameters for the given corpus. The final model along with its parameters are illustrated in figure 6.6. The final number of topics are 9 with a learning rate of 0.05.

```
multi_lda_final_notts = gensim.models.LdaMulticore(corpus = corpus,
                         id2word = id2word,
                         random_state = 42,
                         num_topics = 9,
                         passes=10,
                         chunksize=512,
                         alpha=0.05,
                         offset=64,
                         eta=None,
                         iterations=100,
                         per_word_topics=True,
                         workers=6)
```

Figure 6.6 Final LDA model

```
[(0,
  '0.031*"wait" + 0.026*"double" + 0.020*"thank" + 0.016*"jabbed" + '
  '0.013*"staff" + 0.009*"great" + 0.008*"volunteer" + 0.008*"week" + '
  '0.008*"summer" + 0.007*"look" + 0.007*"queue" + 0.007*"team" + '
  '0.007*"super" + 0.006*"amazing" + 0.006*"soon" + 0.006*"girl" + '
  '0.006*"finally" + 0.006*"minute" + 0.006*"long" + 0.005*"announce"'),
 (1,
  '0.040*"mask" + 0.028*"wear" + 0.010*"public" + 0.010*"continue" + '
  '0.009*"2021" + 0.009*"follow" + 0.008*"july" + 0.007*"state" + 0.007*"stay" '
  '+ 0.006*"news" + 0.006*"student" + 0.006*"drive" + 0.006*"slot" + '
  '0.005*"require" + 0.005*"social" + 0.005*"protect" + 0.005*"indoor" + '
  '0.005*"officially" + 0.005*"order" + 0.005*"service"'),
 (2,
  '0.023*"week" + 0.023*"book" + 0.021*"appointment" + 0.021*"clinic" + '
  '0.017*"july" + 0.017*"centre" + 0.015*"receive" + 0.013*"visit" + '
  '0.013*"available" + 0.012*"walkin" + 0.011*"weekend" + 0.010*"walk" + '
  '0.010*"find" + 0.010*"tomorrow" + 0.009*"come" + 0.009*"open" + '
  '0.008*"site" + 0.008*"free" + 0.007*"need" + 0.006*"near"'),
 (3,
  '0.028*"case" + 0.025*"variant" + 0.020*"death" + 0.017*"delta" + '
  '0.014*"rate" + 0.010*"disease" + 0.009*"number" + 0.009*"effective" + '
  '0.009*"study" + 0.009*"infection" + 0.009*"population" + 0.009*"high" + '
  '0.009*"report" + 0.008*"immunity" + 0.007*"spread" + 0.007*"unvaccinate" + '
  '0.007*"risk" + 0.007*"datum" + 0.007*"protect" + 0.006*"virus"'),
 (4,
  '0.017*"work" + 0.011*"help" + 0.009*"worker" + 0.008*"great" + '
  '0.008*"government" + 0.008*"support" + 0.008*"read" + 0.006*"team" + '
  '0.006*"child" + 0.006*"doctor" + 0.005*"healthcare" + 0.005*"school" + '
  '0.005*"citizen" + 0.005*"ready" + 0.005*"thank" + 0.005*"year" + '
  '0.005*"nurse" + 0.005*"hospital" + 0.005*"drive" + 0.005*"trust"'),
 (5,
  '0.034*"life" + 0.030*"world" + 0.030*"family" + 0.026*"million" + '
  '0.026*"save" + 0.025*"protect" + 0.024*"care" + 0.022*"finally" + '
  '0.016*"medical" + 0.015*"friend" + 0.014*"love" + 0.013*"rest" + '
  '0.013*"vaxxe" + 0.012*"good" + 0.011*"illness" + 0.011*"common" + '
  '0.010*"community" + 0.010*"help" + 0.008*"country" + 0.008*"possible"'),
 (6,
  '0.015*"moderna" + 0.014*"complete" + 0.012*"happy" + 0.011*"question" + '
  '0.009*"fight" + 0.009*"watch" + 0.008*"video" + 0.008*"door" + '
  '0.008*"answer" + 0.008*"covishield" + 0.007*"covaxin" + 0.007*"canada" + '
  '0.006*"yesterday" + 0.006*"news" + 0.006*"make" + 0.006*"england" + '
  '0.005*"feel" + 0.005*"receive" + 0.005*"talk" + 0.005*"great"'),
 (7,
  '0.018*"feel" + 0.013*"virus" + 0.012*"risk" + 0.012*"test" + 0.012*"think" '
  '+ 0.010*"long" + 0.009*"choice" + 0.009*"effect" + 0.008*"positive" + '
  '0.007*"stop" + 0.007*"kill" + 0.007*"believe" + 0.007*"body" + 0.007*"sure" '
  '+ 0.007*"love" + 0.006*"death" + 0.006*"cause" + 0.006*"right" + '
  '0.006*"come" + 0.005*"work"'),
 (8,
  '0.017*"effect" + 0.013*"feel" + 0.011*"post" + 0.008*"happen" + '
  '0.008*"right" + 0.007*"hour" + 0.007*"tweet" + 0.007*"pain" + 0.006*"hope" '
  '+ 0.006*"reaction" + 0.006*"thing" + 0.006*"month" + 0.006*"yesterday" + '
  '0.006*"fine" + 0.005*"true" + 0.005*"sore" + 0.005*"medium" + 0.005*"worry" '
  '+ 0.005*"year" + 0.005*"nice"')]
```

Figure 6.7 Topics from LDA model

Figure 6.7 illustrates the 9 topics created by training the corpus. Each topic is described with the topic it is associated with along with its probability. Each topics says something different. This is derived from the extracted data from twitter. The topics can be seen as a collection of words along with their probability on belonging to the topic. These words help us arrive at a topic for the given words. The topic names for each of the nine topics are illustrated in the figure 6.8. These topics help us summarize the whole of data into these topics.

```
for tpc_ in text_df["Topic_Num"].values.tolist():
    if tpc_ == 0:
        tpc_nms.append("Thank staff for double jab")
    if tpc_ == 1:
        tpc_nms.append("In 2021, continue to wear a mask in public")
    if tpc_ == 2:
        tpc_nms.append("Book appointments for vaccination in clinics or centres")
    if tpc_ == 3:
        tpc_nms.append("Delta variant cases lead to higher death rate")
    if tpc_ == 4:
        tpc_nms.append("Help and support healthcare workers")
    if tpc_ == 5:
        tpc_nms.append("Save the world by protecting families")
    if tpc_ == 6:
        tpc_nms.append("Take your vaccine and fight the disease")
    if tpc_ == 7:
        tpc_nms.append("Stay positive and test regularly")
    if tpc_ == 8:
        tpc_nms.append("Post vaccination there maybe side effects")
```

Figure 6.8 Topics from LDA

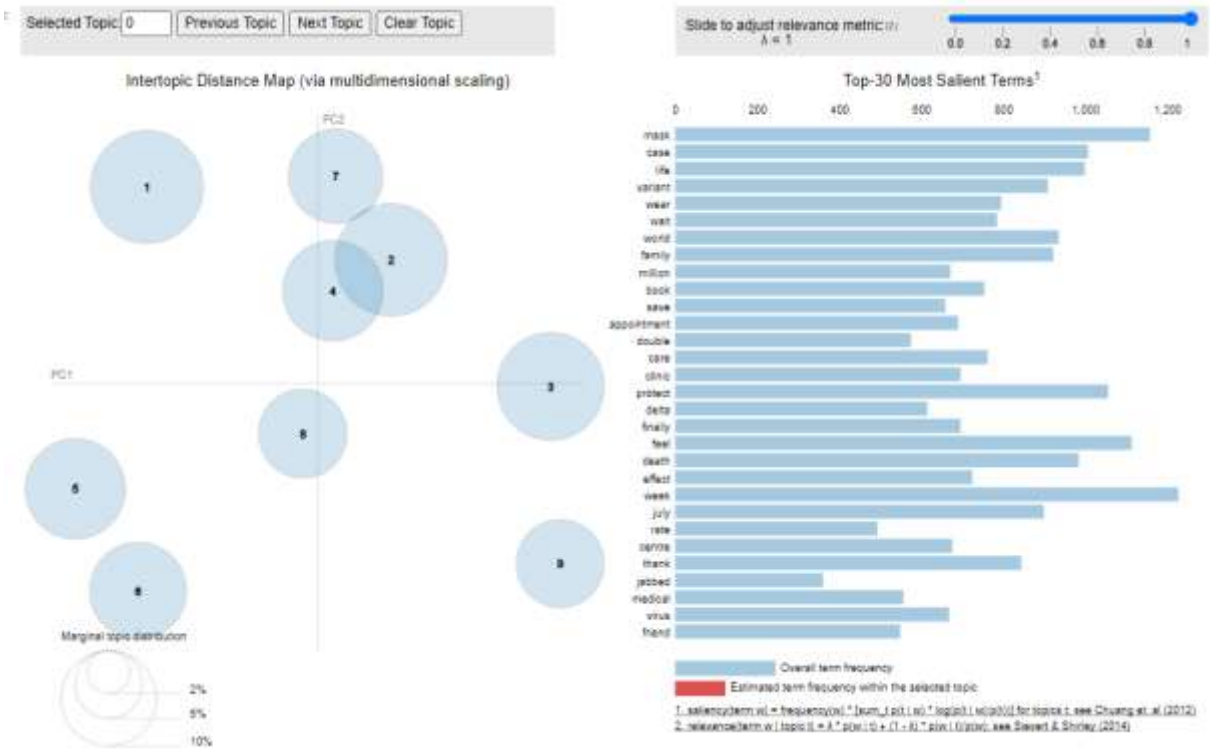

Figure 6.9 Interactive topic modelling

Figure 6.9 illustrates the interactive visualization for topic modelling. This graph is interactive on the ipynb file. It is also available in the html format in the GitHub repository. The topics are represented in a two-dimensional plot. In the left of the graph is the inter topic distance. The topics are numbered from 1 to 9. The size of the bubble represents how frequently the topics occur throughout the documents. An overlapping bubble indicates that there

are common words in both the topics. The amount of overlap indicates the number of common terms in the topics. We see there is a big overlap between topics 4 and 2. There is a slight overlap between 2 and 7. The distance between the bubbles explain the semantic relation between them. Semantic relation means the similarity of topics. If a bubble is closer to a bubble, it is likely that the topics are similar based on their distance. Topics 5 and 3 are very far away, it implies they are talking about very different topics. On the other hand, topics 5 and 6 are close by, its likely that they are like some extent. The horizontal graphs on the right give the frequency distribution of the words in the documents. Figure 6.10 illustrates the frequency distribution of the words a selected topic.

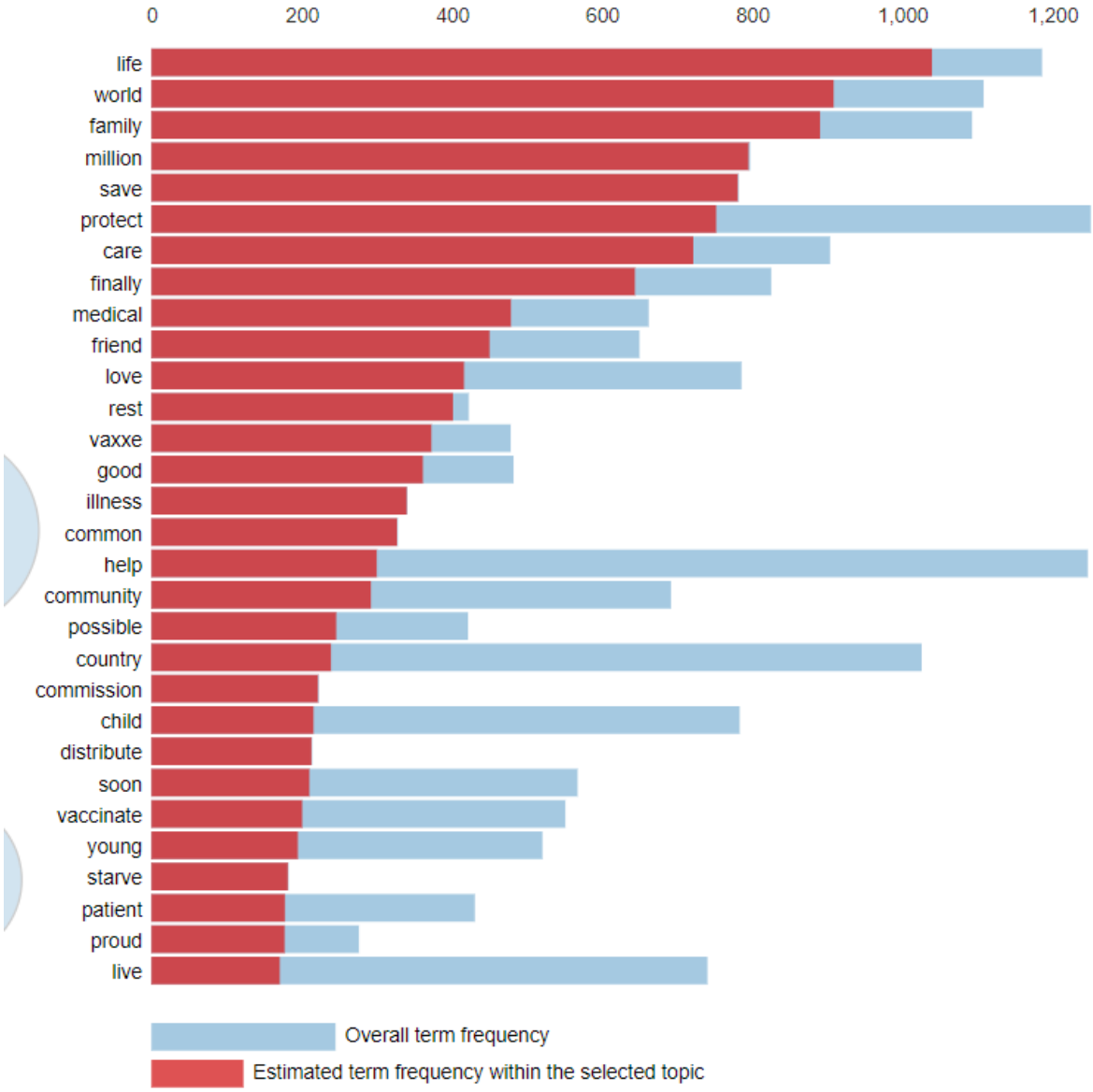

Figure 6.10 The Frequency of each word in a topic

The LDA modelling has given an idea on what topics and words the extracted data is talking about. This is very useful in understanding the public sentiment and opinion. We talked on topic modelling using the data extracted from twitter. The topic modelling gave us the 9 topics and the words along with its probability. This helps us understand the words being written down on tweets. We also visualized the words along with the topics to see the topics through graphs. These graphs also provided us with semantic relations and similarity between the topics.

Chapter 7

Text Classification

7.1 Introduction

In this chapter we will talk about text classification where we will classify the tweets into labels. We will select one of the topics of supreme importance to our discussion. The tweet of this topic is flagged. Since the tweets related to this this topic are very less, we will oversample the data. We split the data into test and train. A pipeline was created to perform tfidf on the data. SGD classifier was used to train and test on the data. We evaluate the model based on various evaluation metrics.

7.2 Text Classification

7.2.1 Text Classification: Introduction

Binary classification is the technique used to classify documents either belonging to a class or not. Topic 6 in the LDA model is of our interest. It is illustrated in figure 7.1. This topic is selected as we want classify tweets on

```
(6,
 '0.015*"moderna" + 0.014*"complete" + 0.012*"happy" + 0.012*"question" + '
 '0.009*"fight" + 0.009*"watch" + 0.008*"video" + 0.008*"door" + '
 '0.008*"answer" + 0.008*"news" + 0.008*"covishield" + 0.007*"covaxin" + '
 '0.007*"canada" + 0.006*"england" + 0.006*"great" + 0.006*"yesterday" + '
 '0.006*"receive" + 0.005*"make" + 0.005*"talk" + 0.005*"feel"'),
```

Figure 7.1 Topic 6

vaccination. After selecting the topic of interest, the tweets are flagged. This dataset is segregated into train and test sets having 25% of test set. The tweets belonging to topic 6 were 3568. It leads to imbalance in the data hence, oversampling is done. A pipeline is created as illustrated in 7.2.

```
pipeline = Pipeline([
    ('vect',TfidfVectorizer(max_df = 0.6, min_df=0.0001, norm = 'l2', use_idf = True, ngram_range = (1,2), max_features = None)),
    ('clf',SGDClassifier(random_state=0,  alpha = 2e-05, penalty = 'l2', loss = 'modified_huber', max_iter = 10))
])
```

Figure 7.2 Classification Pipeline

This pipeline consists of two steps. The first is the tfidf vectorizer and the next is the SGD classifier. We fit the data on the pipeline.

Chapter 6

Conclusion

The main aim of this project was to produce a system which can identify patients at risk of developing target conditions. In summary, this was achieved for a number of conditions. Further, the results from the ANN did improve on that of Jonathan Turner's PhD work [44] and Riccardo Miotto's `Deep patient' paper [26]. However, the number of available EHRs restricted the ability to achieve better results.

I hope that providing this work will benet the future of identifying patients at risk of developing non-communicable conditions. Despite successfully calculating raised condition likelihoods, this alone does not qualify whether the results can be applied in a clinical context. In order to test this, clinical trails would have to be conducted [19] and ethical approval would be required.

I believe this is currently the only open source software which calculates raised condition likelihoods for patients. With the further improvement of technology such as this I believe that personalised medicine will be available for a greater number of people at a reduced cost.

Bibliography

[1] The Visual and Data Journalism Team 30 July 2021, https://www.bbc.co.uk/news/world-51235105

[2] Richard Partington 10 May 2021, https://www.theguardian.com/business/2021/may/10/uk-economy-to-suffer-700bn-output-loss-due-to-covid-and-brexit-thinktank-warns

[3] NHS UK 29 July 2021, https://coronavirus.data.gov.uk/details/healthcare

[4] ABPI UK, https://www.abpi.org.uk/medicine-discovery/covid-19/what-are-pharmaceutical-companies-doing-to-tackle-the-disease/

[5] Michelle Roberts 8 Jan 2021, https://www.bbc.co.uk/news/health-55586410

[6] Josh Holder 30 July 2021, https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html

[7] BBC 8 Dec 2021, https://www.bbc.co.uk/news/uk-55227325

[8] Bing Liu, https://www.kellogg.northwestern.edu/departments/marketing/~/media/06B0845A9D844EDF905BD83871CE5CAA.ashx

[9] Owen Jarus 20 March 2021, https://www.livescience.com/worst-epidemics-and-pandemics-in-history.html

[10] Lois Zoppi 2021, https://www.news-medical.net/health/How-does-the-COVID-19-Pandemic-Compare-to-Other-Pandemics.aspx

[11] CDC 27 May 2021, https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/how-they-work.html

[12] Ewen Callaway, Heidi Ledford, Giuliana Viglione, Traci Watson, & Alexandra Witze 14 Dec 2021, https://www.nature.com/immersive/d41586-020-03437-4/index.html

[13] A.H. Alamoodi, B.B. Zaidan, A.A. Zaidan, O.S. Albahri, K.I. Mohammed, R.Q. Malik, E.M. Almahdi, M.A. Chyad, Z. Tareq, A.S. Albahri, Hamsa Hameed, Musaab Alaa, Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review, Expert Systems with Applications

[14] Xingchen Pan, David M. Ojcius, Tianyue Gao, Zhongsheng Li, Chunhua Pan, Chungen Pan, Lessons learned from the 2019-nCoV epidemic on prevention of future infectious diseases, Microbes and Infection

[15] Kim KS, Sin SC, Yoo-Lee EY. Undergraduates' use of social media as information sources. College & research libraries. 2014 Jul 1;75(4):442-57.

[16] Appel, G., Grewal, L., Hadi, R. et al. The future of social media in marketing. J. of the Acad. Mark. Sci. 48, 79–95 (2020). https://doi.org/10.1007/s11747-019-00695-1

[17] L. DeNardis, A.M. Hackl, Internet governance by social media platforms,Telecommunications Policy

[18] Ali, Kashif & Dong, Hai & Bouguettaya, Athman & Erradi, Abdelkarim & Hadjidj, Rachid. (2017). Sentiment Analysis as a Service: A social media Based Sentiment Analysis Framework. 10.1109/ICWS.2017.79.

[19] Statistica 2019, https://www.statista.com/statistics/579411/top-us-social-networking-apps-ranked-by-session-length/

[20] Kate Crawford (2009) Following you: Disciplines of listening in social media, Continuum, 23:4, 525-535, DOI: 10.1080/10304310903003270

[21] Jansen, B. J., Sobel, K., & Cook, G. (2010). Gen X and Ys attitudes on using social media platforms for opinion sharing. In CHI'10 extended abstracts on human factors in computing systems (pp. 3853–3858).

[22] Chung, W., He, S., & Zeng, D. (2015). emood: Modeling emotion for social media analytics on Ebola disease outbreak.

[23] Ji X., Chun S.A., Geller J. (2016) Knowledge-Based Tweet Classification for Disease Sentiment Monitoring. In: Pedrycz W., Chen SM. (eds) Sentiment Analysis and Ontology Engineering. Studies in Computational Intelligence, vol 639. Springer, Cham. https://doi.org/10.1007/978-3-319-30319-2_17

[24] Sungwoon Choi, Jangho Lee, Min-Gyu Kang, Hyeyoung Min, Yoon-Seok Chang, Sungroh Yoon, Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks

[25] Statistica 2019, https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[26] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010).

[27] Poria, S., Cambria, E., & Gelbukh, A. (2015). Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (pp. 2539-2544).

[28] Dos Santos, C. N., & Gatti, M. (2014, August). Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts

[29] Amananandrai, 04 June 2020, https://dev.to/amananandrai/recent-advances-in-the-field-of-nlp-33o1

[30] Blei, D. M. (2012). Topic modeling and digital humanities. Retrieved from http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

[31] Lyra, M. (2017). Evaluating topic models. Retrieved from https://pydata.org/berlin2017/schedule/presentation/54/

[32] Blei, D. M., Ng, A., Jordan, M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

[33] Deng, X. l., Tang, Y. Q., & Huang, Y. H. (2015). Opinion mining for emergency case risk analysis in spark based distributed system. In Proceedings of the 1st ACM SIGSPATIAL international workshop on the Use of GIS in emergency management (pp. 1–8).

[34] Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., Schoen, H., Gloor, P, Tarabanis, K. J. I. R. 2013). Understanding the predictive power of social media. Internet Research.

[35] Kim, E. H.-J., Jeong, Y. K., Kim, Y., Kang, K. Y., & Song, M. (2016). Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. Journal of Information Science, 42, 763–781.

[36] Jain, V. K., & Kumar, S. (2018). Effective surveillance and predictive mapping of mosquito-borne diseases using social media. Journal of Computational Science, 25, 406–415.

[37] Baker, Q. B., Shatnawi, F., Rawashdeh, S., Al-Smadi, M., & Jararweh, Y. (2020). Detecting epidemic diseases using sentiment analysis of Arabic Tweets. Journal of Universal Computer Science, 26, 50–70.

[38] Fei-Fei L, Perona P (2005) A bayesian hierarchical model for learning natural scene categories. IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 2, pp 524–531

[39] Jiang S, Qian X, Shen J, Fu Y, Mei T (2015) Author topic model-based collaborative filtering for personalized POI recommendations. IEEE Trans Multimedia 17(6):907–918

[40] Phan, X. H., Nguyen, C. T., Le, D. T., Nguyen, L. M., Horiguchi, S., and Ha, Q. T. (2011). A hidden topic-based framework toward building applications with short web documents. IEEE Trans. Knowl. Data Eng. 23, 961–976. doi: 10.1109/TKDE.2010.27

[41] Cheng, X., Yan, X., Lan, Y., and Guo, J. (2014). Btm: topic modeling over short texts. IEEE Trans. Knowl. Data Eng. 26, 1–1. doi: 10.1109/TKDE.2014.2313872

[42] Xie, P., and Xing, E. P. (2013). "Integrating document clustering and topic modeling," in Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (Bellevue, WA), 694–703.

[43] Hofmann, T. (1999). Probabilistic latent semantic analysis," The 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 50–57. doi: 10.1145/312624.312649

[44] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. J. Am. Soc. Inform. Sci. 41, 391–407. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[45] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. JMLR, 993–1022.

[46] Negara, Edi Surya & Triadi, Dendi. (2019). Social Media Analytics: Data Utilization of Social Media for Research.

[47] S. Kumar, F. Morstatter, and H. Liu. (2013) Twitter Data Analytics.[Online]. Available: www. tweettracker.fulton.asu.edu

[48] Negara, E.S. and Andryani, R., 2018. A Review on Overlapping and Non-Overlapping Community Detection Algorithms for Social NetworkAnalytics. Far East Journal of Electronics and Communications, 18 (1), 1-27

[49] S. Stieglitz and D. Linh. (2013) Social media analytics and politicalcommunication; a social media analytics framework, Social NetworkAnalysis and Mining, vol. 3, no. 4, pp. 12771291.

[50] D. Zeng, H. Chen, R. Lusch, and S.-H. Li. (2010) Socialmedia analyticsand intelligence, Intelligent Systems, IEEE, vol. 25, no. 6, pp. 1316.Global Web Index. (2014) Survei data global web index. [Online].Available: https://www. globalwebindex.net/

[51] R. Brussee and E. t. Hekman. (2015) Social media are highly accessible media.

[52] W. Fan and M. D. Gordon. (2014) The power of social media analytics,Communications of the ACM, vol. 57, no. 6, pp. 7481.

[53] W. Fan, L. Wallace, S. Rich, and Z. Zhang. (2006) Tapping the power oftext mining, Communications of the ACM, vol. 49, no. 9, pp. 7682.

[54] Kharde, V. and Sonawane, P., 2016. Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.

[55] A.Pak and P. Paroubek. „Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010, pp.1320-1326

[56] R. Parikh and M. Movassate, "Sentiment Analysis of User- GeneratedTwitter Updates using Various Classi_cation Techniques",CS224N Final Report, 2009

[57] Go, R. Bhayani, L.Huang. "Twitter Sentiment ClassificationUsing Distant Supervision". Stanford University, Technical Paper,2009

[58] L. Barbosa, J. Feng. "Robust Sentiment Detection on Twitterfrom Biased and Noisy Data". COLING 2010: Poster Volume,pp. 36-44.

[59] Bifet and E. Frank, "Sentiment Knowledge Discovery inTwitter Streaming Data", In Proceedings of the 13th InternationalConference on Discovery Science, Berlin, Germany: Springer,2010, pp. 1-15.

[60] Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011Workshop on Languages in Social Media,2011, pp. 30-38

[61] Dmitry Davidov, Ari Rappoport." Enhanced Sentiment Learning Using Twitter Hashtags and Smileys". Coling 2010: Poster Volumepages 241{249, Beijing, August 2010

[62] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social MediaData", IEEE 14th International Conference on Mobile Data Management,Milan, Italy, June 3 - 6, 2013, pp 91-96, ISBN: 978-1-494673-6068-5, http://doi.ieeecomputersociety.org/10.1109/MDM.2013.

[63] Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategyfor Sentiment Analysis on English Tweets", th InternationalWorkshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland,Aug 23-24 2014, pp 171-175.

[64] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

[65] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using wordnet to measure semantic orientations of adjectives," 2004.

[66] R. Xia, C. Zong, and S. Li, “Ensemble of feature sets and classification algorithms for sentiment classification,” Information Sciences: an International Journal, vol. 181, no. 6, pp. 1138–1152, 2011.

[67] ZhunchenLuo, Miles Osborne, TingWang, An effective approachto tweets opinion retrieval", Springer Journal onWorldWideWeb,Dec 2013, DOI: 10.1007/s11280-013- 0268-7.

[68] H. Saif, Y.He, and H. Alani, “SemanticSentimentAnalysisof Twitter,” Proceeding of the Workshop on Information Extraction and Entity Analytics on Social Media Data. United Kingdom: Knowledge Media Institute, 2011.

[69] R. Prabowo, and M. Thelwall, “Sentiment Analysis:A Combined Approach,” International World Wide Web Conference Committee (IW3C2), 2009. UnitedKingdom:Universityof Wolverhamption.

[70] M. Annett, and G. Kondrak, “A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs,” Conference on web search and web data mining (WSDM). University of Alberia: Department of Computing Science, 2009.

[71] T. Carpenter, and T. Way, “Tracking Sentiment Analysis through Twitter,”. ACM computer survey. Villanova:VillanovaUniversity, 2010.

[72] A.Sharma, and S. Dey, “Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis,” Association for the advancement of Artificial Intelligence, 2012.

[73] P. Goncalves, F. Benevenuto, M. Araujo and M. Cha, “Comparing and Combining Sentiment Analysis Methods”, 2013.

[74] M.Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “ Lexicon-Based Methods for Sentiment Analysis,” Association for Computational Linguistics, 2011.

[75] P. Goncalves, F. Benevenuto, M. Araujo and M. Cha, “Comparing and Combining Sentiment Analysis Methods”, 2013.

[76] A. Blom and S. Thorsen, “Automatic Twitter replies with Python,” International conference “Dialog 2012”.

[77] B. Pang, and L. Lee, “Opinion mining and sentiment analysis,” 2nd workshop on making sense of Microposts. Ithaca: Cornell University. Vol.2(1), 2008.

[78] M. Hu, and B. Liu, “Mining and summarizing customer reviews,” 2004.

[79] Sarlan, Aliza & Nadam, Chayanit & Basri, Shuib. (2014). Twitter sentiment analysis. 212-216. 10.1109/ICIMU.2014.7066632.

[80] Yaakub, Mohd Ridzwan & Abu Latiffi, Muhammad Iqbal & Safra, Liyana. (2019). A Review on Sentiment Analysis Techniques and Applications. IOP Conference Series: Materials Science and Engineering. 551. 012070. 10.1088/1757-899X/551/1/012070.

[81] ianqiang, Z., Xiaolin, G., Xuejun, Z.: Deep convolution neural networks for twitter sentimentanalysis, pp. 23253–23260. IEEE Access (2018)

[82] Sumit, S.H., Hossan M. Z., Muntasir, T.A., Sourov T.: Exploring word embedding for Banglasentiment analysis. In: International Conference on Bangla Speech and Language Processing(ICBSLP). IEEE (2018)

[83] Sharma, Dipti & Sabharwal, Munish & Goyal, Vinay & Vij, Mohit. (2020). Sentiment Analysis Techniques for Social Media Data: A Review. 10.1007/978-981-15-0029-9_7.

[84] Pirri, S.; Lorenzoni, V.; Andreozzi, G.; Mosca, M.; Turchetti, G. Topic Modeling and User Network Analysis on Twitter during World Lupus Awareness Day. Int. J. Environ. Res. Public Health 2020, 17, 5440. https://doi.org/10.3390/ijerph17155440

[85] Weng, L.; Menczer, F.; Ahn, Y.-Y. Virality prediction and community structure in social networks. Sci. Rep. 2013, 3, 2522.

[86] Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. Science 2018, 359, 1146–1151.

[87] Paul, M.J.; Dredze, M.; Broniatowski, D. Twitter Improves Influenza Forecasting. PLoS Curr. 2014, 6.

[88] Smolinski, M.S.; Crawley, A.W.; Baltrusaitis, K.; Chunara, R.; Olsen, J.M.; Wójcik, O.; Santillana, M.; Nguyen, A.; Brownstein, J.S. Flu Near You: Crowdsourced Symptom Reporting Spanning 2 Influenza Seasons. Am. J. Public Health 2015, 105, 2124–2130.

[89] Oliver, J.E.; Wood, T. Medical Conspiracy Theories and Health Behaviors in the United States. JAMA Intern. Med. 2014, 174, 817–818.

[90] Miah, S.J.; Hasan, N.; Hasan, R.; Gammack, J. Healthcare support for underserved communities using a mobile social media platform. Inf. Syst. 2017, 66, 1–12.

[91] Thomas, M.; Narayan, P. The Role of Participatory Communication in Tracking Unreported Reproductive Tract Issues in Marginalized Communities. Inf. Technol. Dev. 2016, 22, 117–133.

[92] Young, S.D.; Rivers, C.; Lewis, B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev. Med. (Baltim) 2014, 63, 112–115.

[93] Golder, V.; Morand, E.F.; Hoi, A.Y. Quality of Care for Systemic Lupus Erythematosus: Mind the Knowledge Gap. J. Rheumatol. 2017, 44, 271–278.

[94] Stockl, A. Complex syndromes, ambivalent diagnosis, and existential uncertainty: The case of Systemic Lupus Erythematosus (SLE). Soc. Sci. Med. 2007, 65, 1549–1559.

[95] Gergianaki, I.; Bertsias, G. Systemic Lupus Erythematosus in Primary Care: An Update and Practical Messages for the General Practitioner. Front. Med. 2018, 5, 161.

[96] Reuter, K.; Danve, A.; Deodhar, A. Harnessing the power of social media: How can it help in axial spondyloarthritis research? Curr. Opin. Rheumatol. 2019, 31, 321–328.

[97] Crowe, A.L.; McKnight, A.J.; McAneney, H. Communication Needs for Individuals with Rare Diseases Within and Around the Healthcare System of Northern Ireland. Front. Public Health 2019, 7, 236.

[98] Tenderich, A.; Tenderich, B.; Barton, T.; Richards, S.E.What Are PWDs (PeopleWith Diabetes) Doing Online? A Netnographic Analysis. J. Diabetes Sci. Technol. 2019, 13, 187–197.

[99] Rathore, A.K.; Kar, A.K.; Ilavarasan, P.V. Social Media Analytics: Literature Review and Directions for Future Research. Decis. Anal. 2017, 14, 229–249.

[100] Mao, J.J.; Chung, A.; Benton, A.; Hill, S.; Ungar, L.; Leonard, C.E.; Hennessy, S.; Holmes, J.H. Online discussion of drug side effects and discontinuation among breast cancer survivors. Pharmacoepidemiol. Drug Saf. 2013, 22, 256–262.

[101] Backa, K.E.; Holmberg, K.; Ek, S. Communicating diabetes and diets on Twitter—A semantic content nalysis.Int. J. Netw. Virtual. Organ. 2016, 16, 8–24.

[102] Xu,W.W.; Chiu, I.-H.; Chen, Y.; Mukherjee, T. Twitter hashtags for health: Applying network and content analyses to understand the health knowledge sharing in a Twitter-based community of practice. Qual. uant. 2015, 49, 1361–1380.

[103] Smith, K.T. Hospital Marketing and Communications Via Social Media. Serv. Mark. Q. 2017, 38, 187–01.

[104] Blei, D.M. Probabilistic topic models. Commun. ACM 2012, 55, 77–84.

[105] Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet allocation. J. Mach. Learn. Res. 2003, 3, 993–1022.

[106] Roberts, M.E.; Stewart, B.M.; Tingley, D.; Lucas, C.; Leder-Luis, J.; Gadarian, S.K.; Albertson, B.; and, D.G. Structural topic models for open-ended survey responses. Am. J. Pol. Sci. 2014, 58, 1064–1082.

[107] Roberts, M.E.; Stewart, B.M.; Tingley, D. Stm: An R package for structural topic models. J. Stat. oftw. 2019, 91.

[108] Mimno, D.; Wallach, H.M.; Talley, E.; Leenders, M.; McCallum, A. Optimizing semantic coherence in topic models. In Proceedings of the EMNLP 2011—Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, 27–31 July 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 262–272.

[109] Airoldi, E.M.; Bischof, J.M. Improving and Evaluating Topic Models and Other Models of Text. J. Am. tat. Assoc. 2016, 111, 1381–1403.

[110] Wheeler, L.M.; Pakozdi, A.; Rajakariar, R.; Lewis, M.; Cove-Smith, A.; Pyne, D. 139 Moving with the Times: Social Media Use Amongst Lupus Patients. Rheumatology 2018, 57, key075-363.

[111] Jiang, S. Functional interactivity in social media: An examination of Chinese health care organizations' microblog profiles. Health Promot. Int. 2019, 34, 38–46.

[112] Wang, Y.; McKee, M.; Torbica, A.; Stuckler, D. Systematic Literature Review on the Spread of Health-related Misinformation on Social Media. Soc. Sci. Med. 2019, 240, 112552..

[113] Avasthi, Sandhya. (2020). Topic Modeling on Twitter Data and Identifying Health-Related Issues. 10.1007/978-981-15-4936-6_6.

[114] Jordan, S., Hovet, S., Fung, I., Liang, H., Fu, K. W., & Tse, Z. (2019). Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response. Data, 4(1), 6.

[115] Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. International journal of information management, 39, 156-168.

[116] Paul, M. J., & Dredze, M. (2014). Discovering health topics in social media using topic models. PloS one, 9(8), e103408.

[117] Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011, March). Identifying health-related topics on twitter. In International conference on social computing, behavioral-cultural modeling, and prediction (pp. 18-25). Springer, Berlin, Heidelberg.

[118] Beykikhoshk, A., Arandjelović, O., Phung, D., Venkatesh, S., & Caelli, T. (2015). Using Twitter to learn about the autism community. Social Network Analysis and Mining, 5(1), 22.

[119] Culotta, A. (2010, July). Towards detecting influenza epidemics by analyzing Twitter messages. In Proceedings of the first workshop on social media analytics (pp. 115-122). acm.

[120] Culotta, A. (2013). Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. Language resources and evaluation, 47(1), 217-238.

[121] Kalyanam, J., Katsuki, T., Lanckriet, G. R., & Mackey, T. K. (2017). Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twitter sphere using unsupervised machine learning. Addictive behaviors, 65, 289-295.

[122] Bosley, J. C., Zhao, N. W., Hill, S., Shofer, F. S., Asch, D. A., Becker, L. B., & Merchant, R. M. (2013). Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. Resuscitation, 84(2), 206-212.

[123] Paul, M. J., Sarker, A., Brownstein, J. S., Nikfarjam, A., Scotch, M., Smith, K. L., & Gonzalez, G. (2016). Social media mining for public health monitoring and surveillance. In Biocomputing 2016: Proceedings of the Pacific symposium(pp. 468-479).

[124] Mohan, P., Lando, H. A., & Panneer, S. (2018). Assessment of tobacco consumption and control in India. Indian Journal of Clinical Medicine, 9, 1179916118759289.

[125] Nazar, G. P., Chang, K. C., Srivastava, S., Pearce, N., Karan, A., & Millett, C. (2019). Impact of India's National Tobacco Control Programme on bidi and cigarette consumption: a difference-in-differences analysis. Tobacco control, tobaccocontrol-2018.

[126] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine learning research3.Jan (2003): 993-1022.

[127] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), 15169-15211.Applications, 78(11), 15169-15211.

[128] Chemudugunta, C., Smyth, P., & Steyvers, M. (2007). Modeling general and specific aspects of documents with a probabilistic topic model. In Advances in neural information processing systems (pp. 241-248).

[129] Paul, M. J. (2012, July). Mixed membership Markov models for unsupervised conversation modeling. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 94-104). Association for Computational Linguistics.

[130] Asghari, Mohsen & Sierra-Sosa, Daniel & Elmaghraby, Adel. (2018). Trends on Health in Social Media: Analysis using Twitter Topic Modeling. 558-563. 10.1109/ISSPIT.2018.8642679.

[131] Chapman, B. E., Lee, S., Kang, H. P., & Chapman, W. W. (2011). Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. Journal of biomedical informatics, 44(5), 728-737.

[132] Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: a good place to detect health conditions. PloS one, 9(1), e86191.

[133] Prier, K. W., Smith, M. S., Giraud-Carrier, C., & Hanson, C. L. (2011, March). Identifying health-related topics on twitter. In International conference on social computing, behavioral-cultural modeling, and prediction (pp. 18-25). Springer, Berlin, Heidelberg.

[134] Greaves, F., Laverty, A. A., Cano, D. R., Moilanen, K., Pulman, S., Darzi, A., & Millett, C. Tweets about hospital quality: a mixed methods study. BMJ Qual Saf. 2014 Oct; 23 (10): 838–46. doi: 10.1136/bmjqs2014-002875.

[135] Hawkins, J. B., Brownstein, J. S., Tuli, G., Runels, T., Broecker, K., Nsoesie, E. O., ... & Greaves, F. (2015). Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Saf, bmjqs2015.

[136] Lyu, Joanne Chen et al. "COVID-19 Vaccine-Related Discussion on Twitter: Topic Modeling and Sentiment Analysis." Journal of medical Internet research vol. 23,6 e24435. 29 Jun. 2021, doi:10.2196/24435

[137] Le TT, Cramer JP, Chen R, Mayhew S. Evolution of the COVID-19 vaccine development landscape. Nat Rev Drug Discov. 2020 Oct 04;19(10):667–668. doi: 10.1038/d41573-020-00151-8.

[138] Gottlieb S. America needs to win the coronavirus vaccine race. The Wall Street Journal. 2020. Apr 26, [2021-04-09].

[139] Andre FE, Booy R, Bock HL, Clemens J, Datta SK, John TJ, Lee BW, Lolekha S, Peltola H, Ruff TA, Santosham M, Schmitt HJ. Vaccination greatly reduces disease, disability, death and inequity worldwide. Bull World Health Organ. 2008 Feb;86(2):140–6. doi: 10.2471/blt.07.040089.

[140] Abbasi J. COVID-19 conspiracies and beyond: how physicians can deal with patients' misinformation. JAMA. 2021 Jan 19;325(3):208–210. doi: 10.1001/jama.2020.22018.

[141] Ball P. Anti-vaccine movement could undermine efforts to end coronavirus pandemic, researchers warn. Nature. 2020 May 13;581(7808):251–251. doi: 10.1038/d41586-020-01423-4.

[142] Kwok KO, Lai F, Wei WI, Wong SYS, Tang JW. Herd immunity - estimating the level required to halt the COVID-19 epidemics in affected countries. J Infect. 2020 Jun;80(6):e32–e33. doi: 10.1016/j.jinf.2020.03.027.

[143] Sanche S, Lin YT, Xu C, Romero-Severson E, Hengartner N, Ke R. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. Emerg Infect Dis. 2020 Jul;26(7):1470–1477. doi: 10.3201/eid2607.200282. doi: 10.3201/eid2607.200282.

[144] Tyson A, Johnson C, Funk C. U.S. public now divided over whether to get COVID-19 vaccine. Pew Research Center. 2020. Sep 17, [2021-04-15].

[145] Ashkenazi S, Livni G, Klein A, Kremer N, Havlin A, Berkowitz O. The relationship between parental source of information and knowledge about measles / measles vaccine and vaccine hesitancy. Vaccine. 2020 Oct 27;38(46):7292–7298. doi: 10.1016/j.vaccine.2020.09.044.

[146] Cole-Lewis H, Pugatch J, Sanders A, Varghese A, Posada S, Yun C, Schwarz M, Augustson E. Social listening: a content analysis of e-cigarette discussions on Twitter. J Med Internet Res. 2015 Oct 27;17(10):e243. doi: 10.2196/jmir.4969.

[147] Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: Twitter and antibiotics. Am J Infect Control. 2010 Apr;38(3):182–8. doi: 10.1016/j.ajic.2009.11.004.

[148] Lazard AJ, Scheinfeld E, Bernhardt JM, Wilcox GB, Suran M. Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. Am J Infect Control. 2015 Oct 01;43(10):1109–11. doi: 10.1016/j.ajic.2015.05.025.

[149] Masri S, Jia J, Li C, Zhou G, Lee M, Yan G, Wu J. Use of Twitter data to improve Zika virus surveillance in the United States during the 2016 epidemic. BMC Public Health. 2019 Jun 14;19(1):761. doi: 10.1186/s12889-019-7103-8.

[150] Signorini A, Segre A, Polgreen Philip M. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One. 2011 May 04;6(5):e19467. doi: 10.1371/journal.pone.0019467.

[151] Bonnevie E, Gallegos-Jeffrey A, Goldbarg J, Byrd B, Smyser J. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. J Commun Healthc. 2020 Dec 15;14(1):12–19. doi: 10.1080/17538068.2020.1858222.

[152] Hussain A, Tahir A, Hussain Z, Sheikh Zakariya, Gogate Mandar, Dashtipour Kia, Ali Azhar, Sheikh Aziz. Artificial intelligence-enabled analysis of public attitudes on Facebook and Twitter toward COVID-19 vaccines in the United Kingdom and the United States: observational study. J Med Internet Res. 2021 Apr 05;23(4):e26627. doi: 10.2196/26627.

[153] Deiner MS, Fathy C, Kim J, Niemeyer K, Ramirez D, Ackley SF, Liu F, Lietman TM, Porco TC. Facebook and Twitter vaccine sentiment in response to measles outbreaks. Health Informatics J. 2019 Sep 17;25(3):1116–1132. doi: 10.1177/1460458217740723.

[154] Benis A, Seidmann A, Ashkenazi S. Reasons for taking the COVID-19 vaccine by US social media users. Vaccines (Basel) 2021 Mar 29;9(4):315. doi: 10.3390/vaccines9040315.

[155] Puri N, Coomes EA, Haghbayan H, Gunaratne K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Hum Vaccin Immunother. 2020 Nov 01;16(11):2586–2593. doi: 10.1080/21645515.2020.1780846.

[156] Malik AA, McFadden SM, Elharake J, Omer SB. Determinants of COVID-19 vaccine acceptance in the US. EClinicalMedicine. 2020 Sep;26:100495. doi: 10.1016/j.eclinm.2020.100495.

[157] Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav. 2021 Mar 05;5(3):337–348. doi: 10.1038/s41562-021-01056-1.

[158] Plutchik R. A general psychoevolutionary theory of emotion. In: Robert P, Henry K, editors. Theories of Emotion. Cambridge, MA: Academic Press; 1980. pp. 3–33.

[159] Misuraca M, Alessia F, Germana S, Maria S. Sentiment Analysis for Education with R: packages, methods and practical applications. ArXiv. Preprint posted online on May 08, 2020

[160] Turney P, Mohammad S. Emotions evoked by common words and phrases: using Mechanical Turk to create an emotion lexicon. Proceedings of the NAACL HLT 2010.

[161] Naldi M. A review of sentiment computation methods with R packages. ArXiv. Preprint posted online on January 24, 2019

[162] Akst J. Russia approves world's first coronavirus vaccine. The Scientist. 2020. Aug 11, [2021-06-22].

[163] Twitter, Inc., 2021, https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

[164] Twitter, Inc., 2021, https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-thetwitter-api

[165] Wisdom, Vivek. (2016). An introduction to Twitter Data Analysis in Python. 10.13140/RG.2.2.12803.30243.

[166] Twitter, Inc., 2021, https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview

[167] Wikipedia, https://en.wikipedia.org/wiki/Lemmatisation

[168] Parthvi Shah, 27 Jun 2020, https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524

[169] Bing Liu, 22 April 2012, https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

[170] Jónsson, Elías. "An Evaluation of Topic Modelling Techniques for Twitter." (2016).

[171] Monkey Learn, https://monkeylearn.com/blog/introduction-to-topic-modeling/

[172] Shashank, 19 Aug 2019, https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[173] Signorini A, Segre A, Polgreen Philip M. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One. 2011 May 04;6(5):e19467. doi: 10.1371/journal.pone.0019467.

[174] Bonnevie E, Gallegos-Jeffrey A, Goldbarg J, Byrd B, Smyser J. Quantifying the rise of vaccine opposition on Twitter during the COVID-19 pandemic. J Commun Healthc. 2020 Dec 15;14(1):12–19. doi: 10.1080/17538068.2020.1858222.

[175] Hussain A, Tahir A, Hussain Z, Sheikh Zakariya, Gogate Mandar, Dashtipour Kia, Ali Azhar, Sheikh Aziz. Artificial intelligence-enabled analysis of public attitudes on Facebook and Twitter toward COVID-19 vaccines in the United Kingdom and the United States: observational study. J Med Internet Res. 2021 Apr 05;23(4):e26627. doi: 10.2196/26627.

[176] Deiner MS, Fathy C, Kim J, Niemeyer K, Ramirez D, Ackley SF, Liu F, Lietman TM, Porco TC. Facebook and Twitter vaccine sentiment in response to measles outbreaks. Health Informatics J. 2019 Sep 17;25(3):1116–1132. doi: 10.1177/1460458217740723.

[177] Benis A, Seidmann A, Ashkenazi S. Reasons for taking the COVID-19 vaccine by US social media users. Vaccines (Basel) 2021 Mar 29;9(4):315. doi: 10.3390/vaccines9040315.

[178] Puri N, Coomes EA, Haghbayan H, Gunaratne K. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. Hum Vaccin Immunother. 2020 Nov 01;16(11):2586–2593. doi: 10.1080/21645515.2020.1780846.

[179] Malik AA, McFadden SM, Elharake J, Omer SB. Determinants of COVID-19 vaccine acceptance in the US. EClinicalMedicine. 2020 Sep;26:100495. doi: 10.1016/j.eclinm.2020.100495.

[180] Loomba S, de Figueiredo A, Piatek SJ, de Graaf K, Larson HJ. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. Nat Hum Behav. 2021 Mar 05;5(3):337–348. doi: 10.1038/s41562-021-01056-1.

[181] Plutchik R. A general psychoevolutionary theory of emotion. In: Robert P, Henry K, editors. Theories of Emotion. Cambridge, MA: Academic Press; 1980. pp. 3–33.

[182] Misuraca M, Alessia F, Germana S, Maria S. Sentiment Analysis for Education with R: packages, methods and practical applications. ArXiv. Preprint posted online on May 08, 2020

[183] Turney P, Mohammad S. Emotions evoked by common words and phrases: using Mechanical Turk to create an emotion lexicon. Proceedings of the NAACL HLT 2010.

[184] Naldi M. A review of sentiment computation methods with R packages. ArXiv. Preprint posted online on January 24, 2019

[185] Akst J. Russia approves world's first coronavirus vaccine. The Scientist. 2020. Aug 11, [2021-06-22].

[186] Twitter, Inc., 2021, https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

[187] Twitter, Inc., 2021, https://developer.twitter.com/en/docs/twitter-api/getting-started/getting-access-to-thetwitter-api

[188] Wisdom, Vivek. (2016). An introduction to Twitter Data Analysis in Python. 10.13140/RG.2.2.12803.30243.

[189] Twitter, Inc., 2021, https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview

[190] Wikipedia, https://en.wikipedia.org/wiki/Lemmatisation

[191] Parthvi Shah, 27 Jun 2020, https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524

[192] Bing Liu, 22 April 2012, https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf

[193] Jónsson, Elías. "An Evaluation of Topic Modelling Techniques for Twitter." (2016).

[194] Monkey Learn, https://monkeylearn.com/blog/introduction-to-topic-modeling/

[195] Shashank 19 Aug 2019, https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Appendix A

Schemas

There are four distinct schemas used in the BigQuery tables:

1. Practice Fusion condition schema (A.1)
2. Practice Fusion patient schema (A.2)
3. Synthea condition schema (A.3)
4. Synthea patient schema (A.4)

Field name	Type	Mode	Description
DiagnosisGuid	STRING	NULLABLE	Primary key for the table
PatientGuid	STRING	NULLABLE	The ID of the patient
ICD9Code	STRING	NULLABLE	The diagnosis code
DiagnosisDescription	STRING	NULLABLE	Human readable name for the diagnosis
StartYear	INTEGER	NULLABLE	The date of diagnosis

Table A.1: The schema for the Practice Fusion condition data

Field name	Type	Mode	Description
PatientGuid	STRING	NULLABLE	The ID of the patient
Sex	STRING	NULLABLE	The sex of the patient
BirthDate	INTEGER	NULLABLE	The patient's birth date

Table A.2: The schema for the Practice Fusion patient data

Field name	Type	Mode	Description
c.code.text	STRING	NULLABLE	The code of the diagnosis
subject.patientId	STRING	NULLABLE	The id of the patient
c.code.coding.system	STRING	NULLABLE	The coding system used

Table A.3: The schema for the Synthea condition data

Field name	Type	Mode	Description
Patient.id	STRING	NULLABLE	The id of the patient
Patient.sex	STRING	NULLABLE	The sex of the patient
Patient.birthDate	STRING	NULLABLE	The patient's birth date

Table A.4: The schema for the Synthea patient data

Appendix B

Execution instructions

To execute the program, FHIR EHR data must be stored in BigQuery using the schemas specied in Appendix A and the table names specied in sql_queries.py. A tutorial on how to upload data to BigQuery is found here[1].

Once data has been uploaded, the Python requirements must be satised. Move into the code/ directory and using Python 3 install the required packages: pip requirements.txt . To run the system, execute the le run_model.sh. The mode of execution depends on the ags specied in main.py.

The full list of ags is listed in Table B.1.

Example output for training and predicting is found in output.txt.

1https://codelabs.developers.google.com/codelabs/cpb200-loading-data/

Flag name Execution	Possible values	Description
modes		
gen_new_seqex	True | False	Whether new seqex should be generated
predict	True | False	Whether predictions should be made
report	True | False	Whether to export reports
hparam_opt	True | False	Enable hyperparameter optimisation
fusion_data	True | False	Whether the fusion data should be used
Directories		
train_dir	"tf_tmp/train.tfrecords"	Species the directory of the training records
eval_dir	"tf_tmp/valid.tfrecords"	Species the directory of the validation records
predict_dir	"tf_tmp/eval.tfrecords"	Species the directory of the evaluation records
model_dir	"model_tmp/"	Species the directory of the model
Mode names		
train_mode	"TRAIN"	Keyword for training
eval_mode	"EVAL"	Keyword for evaluation
train_mode	"PREDICT"	Keyword for prediction
Classier types		
classifier	"DNN" | "Linear"	Species which classier to use
Data management		
project_id	Your Google cloud project id	Species the gcloud project
limit	True | False	Species whether query results should be limited
limit_num	Integer	The maximum number of queries to retrieve
Hyperparamter optimisation		
loss_objective	"auc" | "recall" | "auc_precision_recall"	Species which metric to minimise
target_condition	A diagnosis code relevant to the clinical taxonomy	Species the condition to classify for

Table B.1: The full list of execution ags