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

Tuesday 22nd June 2021

Abstract

A pandemic of covid-19 has claimed almost 3.8 million lives worldwide. There has been a loss of business for many businesses. The economy has suffered because of job losses. Patient numbers soared at the hospitals. Healthcare workers were completely under pressure. In the meantime, pharmaceutical companies and vaccine researchers were seeking a cure for this disease that is spreading at an alarming rate. It did not take long for many companies to discover the vaccines. There were vaccine trials and a study of the patients being conducted. It was Moderna that became available in April 2020. A total of 14 companies marketed vaccines to prevent covid during this time frame. The first vaccine has been received by nearly 20,8% of the global population. Canada, Israel, and the United Kingdom are leading the vaccination race. On 14th December 2020, the first dose was administered. The covid vaccine has now been used for six months, and a lot of people have experienced immunization.

Tweets, or messages posted to Twitter, are short text messages called "tweets" by users on Twitter. Companies can use sentiment analysis on Twitter to take notice of what the voice of their customers is saying about their brand - and their competitors - and to uncover new trends. In sentiment analysis, subjective information is identified and classified in text data. This means that we can determine how we feel about a particular topic or product feature. The process of sentiment analysis uses Natural Language Processing (NLP) and machine learning techniques to automatically interpret human language. An analysis of Twitter data involves five steps: 1. Gather related Twitter data 2. Clean your data using pre-processing techniques 3. Build a machine learning sentiment analysis model 4. Analyse your Twitter data using your sentiment analysis model 5. Visualize your results. A tweet sentiment analysis adds a new dimension to social media monitoring. Analysing tweets in real-time, as well as identifying the sentiment behind each tweet, offers incredible possibilities. Monitoring customer emotions and understanding their feelings on Twitter is one of the most important and powerful tools businesses can use to enhance their social media performance. You can sort data by sentiment manually, but what happens when your data becomes bigger and bigger? Sentiment analysis with machine learning is a fast, scalable, and consistent method that produces accurate results. Regarding COVID-19 vaccines, humans could not possibly read and digest all the tweets.

The Twitter data can be used to determine the sentiment of people about the covid vaccine. Several people have been tweeting their thoughts about the covid vaccine they received over the last six months. Several steps are involved in this process, including data collection, pre-processing of the data, separating positive and negative tweets, and clustering negative tweets. To access Twitter data, you must first sign up for Twitter developer access, which gives you API keys and tokens. The tweepy package in Python allows us to directly extract Twitter data. Extracted data is stored in a database, an excel file, or a CSV file. There are tweets related to Pfizer/BioNTech, Sinopharm, Sinovac, Moderna, Oxford/AstraZeneca, Covaxin, and Sputnik V. An exploratory data analysis is performed to gain an understanding of the tweet data.

Pre-processing is performed on the data. As part of the pipeline, tokens are tokenized, lemmatized, stemmed, and stop words are removed. This bag of words is created by count-vectorization or TF-IDFs. Tweets are rated based on their sentiment to determine if they are positive or negative. Training and testing are conducted with different sets of data. A model is built using machine learning based on the data. A test set is used to validate the predictions. A subset of negative tweets is then extracted from the pre-processed data. Unsupervised machine learning is used to uncover the meaning behind negative tweets. Research relating to public health can then be carried out on the words used in these studies.

Chapter 1

Contextual Background

1.1 Introduction

In the introduction, this chapter explains the motivations for the project and demonstrates its importance. In addition, the project's objectives are outlined.

1.2 Motivations

1.2.1 The Pandemic

Pandemics have emerged and ended many times throughout history, with medical and scientific understanding, living conditions, and socio-political contexts all affecting how they are identified, controlled, and ended. As well as causing different symptoms, certain viruses and infections may affect different populations, and some populations may be more vulnerable than others. That is how pandemics differ from each other. Unlike the 1918 influenza which caused severe symptoms in young and healthy people, COVID-19 mainly affects people over 65 and those with underlying health conditions [1]. An emerging pandemic like COVID-19 can be better understood through accurate case reporting and accessible testing, which helps health organizations and governments curb the spread of the virus in their environments and offer effective advice to people to stop the spread and prevent infection.

The world has become immune to a disease it was not even aware of a year ago, because of a historic scientific effort. Rapid development and testing of COVID-19 vaccines has taken place. Approximately 50 of those drugs are at various stages of clinical testing, according to the WHO in November [2]. SARS-CoV-2 is inactivated in several ways-from crude chemicals to a newer technology never employed in vaccine production. In terms of effectiveness, Pfizer's and Moderna's vaccines may prevent COVID-19 to an extent of 95%, while AstraZeneca's and Oxford's may prove to be less effective [2]. Several significant questions remain, including the effectiveness of vaccines in preventing serious diseases in older people and how long do vaccines provide protection? Vaccination may not prevent the spread of this virus, as many vaccines for other illnesses fail to do so.

1.2.2 Using Twitter as a Data Source

Social media platforms such as Twitter and Facebook offer valuable social data and evidence that has been largely untouched. The internet generates an enormous amount of data every day about a variety of topics, so the internet represents one of the most important sources of information about 21st century society. Several application programming interfaces (APIs) on Twitter make it the most popular academic platform. The use of Twitter data does, however, result in researchers and partners forging a contractual relationship with the organisation. This relationship can have a significant effect on research costs (monetary or time), can impede the technical implementation of research methods and raise ethical and legal issues that should be considered.

1.2.3 Technological Advancements 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) [11]. 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 [11].

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 [11]. 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 [11]. 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 [11]. 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 [11].

1.3 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. Tweets relating to covid vaccine gathered from twitter.
2. Data warehousing is the process of storing data over time.
3. Analysing the data using descriptive statistics.
4. Data is passed into the pre-processing pipeline.
5. Categorizing the tweets according to their positive or negative sentiment.
6. Classifying negative and positive tweets using the machine learning model.
7. Creating clusters of negative tweets.

Chapter 2

Technical Background

2.1 Introduction

The purpose of this chapter is to give the technical background necessary to understand the development of the system Our first step is to look at how we get data through Twitter API. This is followed by data warehousing techniques. Statistical information about the tweets stored in the database. Our next topic of discussion will be the pre-processing pipeline and how it will provide us with clean text data. In the following section, we discuss ways to categorize our data into positive and negative tweets.

Our last step is to create a machine learning model to classify the tweets into negative and positive sentiments.

2.2 Data Collection and Processing

2.2.1 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.

2.2.2 Twitter API

Programmatically accessing and analysing Twitter data can be accomplished using the API, as well as participating in the conversations. Several different resources are accessible through this API, including tweets, users, direct messages, lists, trends, media, and places [3]. A developer account must be applied for, and your use case must be approved before using the Twitter API. If you are interested in academic research, you may apply to the Academic Research product track. Both options offer tailored support, access levels, and pricing [4]. Through the academic research track, academic researchers have access to enhanced functionality, including access to the full-archive search endpoint, a higher tweet limit, and the ability to filter posts with the filtered stream and recent search endpoints [4]. The credentials for this account will be generated once the account has been approved and the Project and App have been created. The Twitter service can be accessed via a REST API. To access twitter's REST API, we are going to use a Python library called tweepy. Twitter REST APIs are easily accessed with its wrapper classes [5]. Our app requires authorization from Twitter to be able to access it. Data from twitter is retrieved through Python using these credentials.

2.2.3 Storing Twitter Data

A MySQL database stores Twitter data. The connection and schema database are setup once MySQL is configured [7]. Data from twitter is gathered into a table. Column names and data types are entered into a table during the creation process. Tweets are returned by all Twitter APIs using JavaScript Object Notation (JSON) [6]. 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. Besides storing the username of the writer of the tweet, the time it was posted, the tweet, the number of retweets, the destination of the tweet and the location, we want to keep track of the information regarding time, place, and location.

2.2.4 Preparing the Data

It is very important to pre-process data before analysing them. The pre-processing of data means taking the data in, preparing it for maximum accuracy by considering our requirements [5]. Thus, understanding the structure of each tweet and analysing its components is needed for pre-processing.

Tweets cannot exceed 140 characters in length. Microblogging's nature (quick and brief messages) leads people to use acronyms or misspell, use emojis or use other characters that convey important messages. An explanation of Twitter terminology is provided below. An emoji is a pictorial representation of a facial expression using letters and punctuation; it normally represents the user's mood. A tweet is referenced by the "@" symbol so that its recipients can be notified. Such a reference automatically alerts the other user. Topics are generally marked with hashtags by users. Basically, they do this so that their tweets are made more visible.

2.2.4.1 Tokenizing the Tweet

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. For @mentions and #hashtags to be correctly tokenized, even the NLTK library requires a few pre-processing steps [5]. Exceptions for hashtags and mentions are based on regular expressions [5]. By tokenizing the text, we prepare it for the next step, which involves removing stop-words like 'the', 'or', 'to', 'and' etc.

2.2.4.2 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.

2.2.5 Analysing the Data

We can then proceed with different analysis objectives after pre-processing and tokenizing the text data.

2.2.5.1 Frequency of Words

Twitter data analysis begins with a simple task of counting the number of occurrences of a term. This allows us to analyse what a particular user frequently tweets about for a particular user. Term frequencies can be used by advertising companies to target ads based on a user's term frequency [5]. Using this method, it is more likely for users to click on a promoted website or visit one.

2.2.5.2 Trending hashtags

Twitter hash tags are among the most widely used features. Historically, they represented world events which were currently taking place. Using hashtags effectively enables us to see a variety of tweets related to the hashtag.

2.2.5.3 Most Used Mentions

People on Twitter are mentioned using '@username' in tweets. As well as replying to someone or notifying multiple people of a particular information, they are used in groups. Based on the followers of a user's most frequent contacts, we can identify which contacts are most likely to mention the user [5].

2.2.6 Feature Extraction

There are several distinctive characteristics of the pre-processed dataset. With the characteristic extraction method, we extract the features from the processed dataset. The polarity of a sentence can later be computed through this aspect, which is useful when determining an individual's opinion using models such as unigrams and bigrams [9]. To process a text or document, machine learning techniques must represent its key attributes [9]. A classification task is accomplished by using these key features, which are called feature vectors.

2.2.6.1 Parts of Speech Tags

A good indicator of subjectivity and sentiment is the use of adjectives, adverbs and some groups of verbs and nouns [9]. Analysing dependency trees or parsing can yield syntactic dependency patterns.

2.2.6.2 Frequencies of Words

Feature models are called monograms, bigrams, and n-grams with their frequency counts. The presence of a word rather than its frequency has more frequently been used to describe this characteristic. The results of Pangeet al. [10] were better when presence was used instead of frequency.

2.2.6.3 Negation

There is a degree of difficulty in interpreting negative aspects of the data. There is usually a polarity shift when there is a negation involved.

2.2.7 Labelling the Data

The Sentiment Analyzer uses NLTK features and classifiers to implement and facilitate Sentiment Analysis tasks, especially for teaching and demonstration purposes. We calculate the polarity, subjectivity, sentiment, negative, positive, neutral, and compound parameters from the cleaned text again [8].

2.2.8 Supervised Learning

The model is built on data with labels, which are then provided to the model during the build process. As a result of this labelling, relevant outputs can be obtained when encountered during decision making. Detecting sentiment with this method requires the selection and extraction of the right set of features. In sentiment analysis, supervised classification is the most prevalent machine learning approach. In machine learning, two sets of data are required: the train set and the test set.

To classify tweets into classes, a variety of machine learning strategies have been developed. Sentiment analysis has enjoyed great success using naive bayes (NB), maximum entropy (ME), and support vector machines (SVM) [9]. Collecting training datasets is the first step in machine learning. After the training data is gathered, we train our classifier. The decision to select a feature is an important step in choosing a supervised classification technique.

2.2.9 Evaluation of Sentiment Classification

Classification involves assigning a class (category) to an object based on previous observations of objects of that category. Classifiers are systems that perform Classification. When a classifier is tested on a set of unknown classes, we can categorize the results in four ways: true positives, false positives, true negatives, and false negatives. Classifiers are used to point out patients who are likely to develop a specified condition.

2.2.9.1 True positives

A true positive is defined when the sentiment correctly classifies themselves as having the target sentiment.

2.2.9.2 False positives

False positives result from the sentiment incorrectly classified as members of the target sentiment.

2.2.9.3 True negatives

True negativity is defined as an emotion that fails to develop the target sentiment, and thus is handled correctly as not developing that emotion.

2.2.9.4 False negatives

A false negative is defined as a sentiment that develops the target sentiment, but which is mistakenly classified as not developing the target sentiment.

2.2.9.5 Equations for Accuracy, Precision, and Recall

The performance of the classifier can be evaluated by four equations.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

$$\text{Precision} = TP/(TP+FP)$$

$$\text{Recall} = TP/(TP+FN)$$

$$F1 = (2\times\text{Precision}\times\text{Recall})/(\text{Precision}+\text{Recall})$$

2.3 Sentiment Analysis

In the comments or tweets, one can find the sentiment, which can be used for a variety of purposes [12]. In addition, the author [13,14] 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 [15].

Most machine learning algorithms are based on supervised classification approaches that receive binary responses representing positive and negative sentiment [16]. To train classifiers, it is necessary to use labelled data [17]. 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 [18]. A strong negative or positive value is negated when it is shifted, which correctly models a mixed perspective. [19] 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 [12].

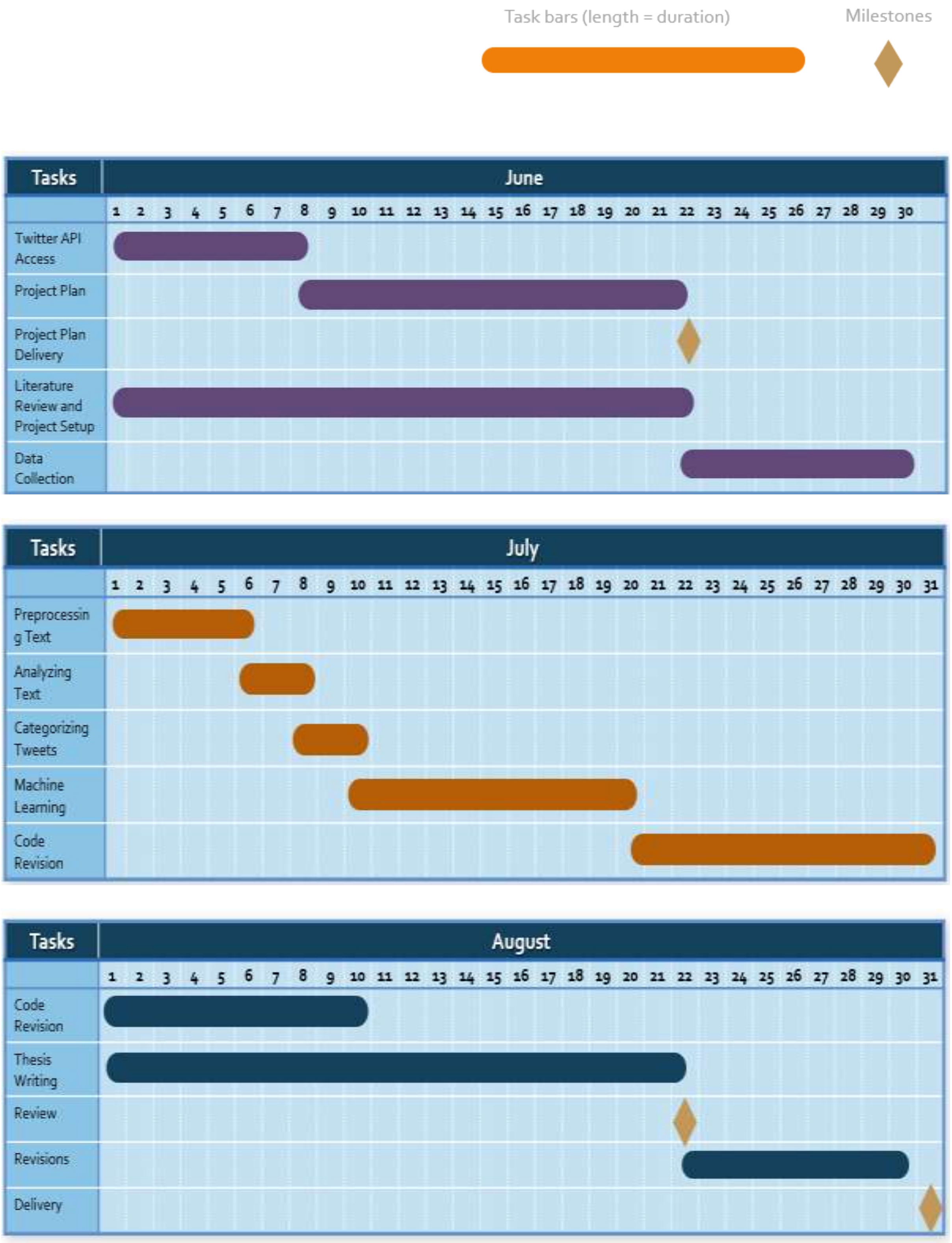
An NLP approach uses machine learning, especially statistical learning, which uses an algorithm that combines a large sample of data with a large corpus to learn the rules [20]. Naturally, Language Processing has been used to provide sentiment analysis at various granularities. Classification began as a document-level classification task [21], and later became a sentence-level classification task [22]. An NLP system is a way to make computers interact with the real world by using human language and input to gather meaning [24].

Generally, Support Vector Machines (SVM), Naive Bayes, and N-Gram are the most popular Machine Learning methods [25]. 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 [25]. Furthermore, the Naive Bayes classifier, despite its simplicity, often outperforms more sophisticated classification methods, despite its simplicity [25]. 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, [25] 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 [26,27]. As related words usually express the same polarity, the semantic relationship must be examined [28]. Researchers use SVM and Naive Bayes to compare alternative approaches to their proposed work [28]. Using these two algorithms provides high accuracy with feature selection. [28].

Appendix A

Project Timeline



Appendix B

Risk Assessment

#	Risk	Description	Probability	Severity	Actions to Minimise Risk
1	**Internal Risks**				
1.1	**Twitter Developer Access**	The Twitter API allows data to be collected through developer access.	**Insignificant**	**Major**	According to Twitter's terms and conditions, developers must use their developer accounts accordingly. Researchers may be put at risk by any actions that violate twitter's terms of service.
1.2	**System Shut Down**	The process of collecting data takes several days to complete. In the event of a shutdown, the data collection process and project delivery may be delayed.	**Moderate**	**Moderate**	It is very important to design the data collection so that shutting down of the system cannot negatively affect it. It is essential to store all the information retrieved from twitter. Backups must be automatically created in multiple locations.
1.3	**Code Files**	Code files without multiple copies can lead to chaos. In case of a catastrophe, the hard drive is wiped out or the system will not start.	**Insignificant**	**Major**	Version control should be used for code files. It is necessary to store multiple copies of the code.

Bibliography

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

[2] Ewen Callaway, Heidi Ledford, Giuliana Viglione, Traci Watson, & Alexandra Witze 14 December 2020, https://www.nature.com/immersive/d41586-020-03437-4/index.html

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

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

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

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

[7] Daniel Foley, 02 October 2018, https://towardsdatascience.com/streaming-twitter-data-into-a-mysql-database-d62a02b050d6

[8] Yalin Yener, 07 November 2020, https://towardsdatascience.com/step-by-step-twitter-sentiment-analysis-in-python-d6f650ade58d

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

[10] Pang, B.and Lee, L. “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”. 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04). 2004, 271-278.

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

[12] 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.

[13] 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.

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

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

[16] 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.

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

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

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

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

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

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

[23] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R.Passonneau, “Sentiment Analysis of Twitter Data,” Annual International Conferences. New York:Columbia University, 2012.

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

[25] 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.

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

[27] 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)

[28] 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.