012 Element – 2022 MOD007893 TRI2 F01CAM

# **Research Project Proposal –** Sentiment Analysis of COVID-19 Vaccines


| | |
|---|---|
| Chukwuemeka JAJA-WACHUKU | SID 2175427 |
| Adam OHINDASA | SID 2169785 |
| Mehmet Huseyin YILDIRIM | SID 2177429 |
| Thi Nhu Ngoc VO | SID 2179088 |

Anglia Ruskin University

Cambridge, United Kingdom – March 2023

# Table of Contents

# Task Allocation

R = Responsible A = Accountable C = Consulted I = Informed

| SID | 2175427 | 2169785 | 2177429 | 2179088 |
|---|---|---|---|---|
| Project requirement & problem analysis | I | I | R&A | I |
| Literature review | I | I | I | R&A |
| Solution design | R&A | I | I | I |
| Technologies & architecture | C&A | R | I | C |
| Evaluation of sentiments | R&A | I | I | I |
| Proposal compilation | I | I | C | R&A |
| Presentation | R | R | R | R&A |

# Abbreviations

| | |
|---|---|
| API | Application Programming Interface |
| AWS | Amazon Web Services |
| COVID-19 | Coronavirus disease 2019 |
| GCP | Google Cloud Platform |
| IE | Information Extraction |
| IR | Information Retrieval |
| Kafka | Apache Kafka |
| ML | Machine Learning |
| NLP | Natural Language Processing |
| NLTK | Natural Language Toolkit |
| Spark | Apache Spark |
| SQL | Structured Query Language |
| VADER | Valence Aware Dictionary and Sentiment Reasoner |

# List of Figures

# Abstract

In December 2019, the COVID-19 pandemic caused by the new coronavirus SARS-CoV-2 erupted suddenly. According to the World Health Organization (*WHO Coronavirus (COVID-19) Dashboard*, 2023), hundreds of millions of verified cases and over 6 million confirmed fatalities have been documented globally. The whole pharmaceutical industry is at war with COVID-19 and is obligated to provide its vaccine to the entire planet as quickly as feasible. Through both traditional and social media, individuals have expressed their sentiments regarding various mass vaccination programs. This project is to examine COVID-19 Vaccine-related tweets and provide a report based on such analysis. The methodology for this project involves collecting tweets related to COVID-19 vaccines, pre-processing the data, performing sentiment analysis on the pre-processed data and storing the sentiment scores in an online database. Finally, sentiment analysis results are visualised using plots and graphs. Expected outcome of the research are identified themes and temporal trends in the sentiment of COVID-19 vaccine-related tweets and explored variations in sentiment.


# 1. Introduction

The COVID-19 pandemic has had a significant impact on populations across the globe, with even developed countries facing challenges in containing its spread despite the implementation of stringent measures. The rapid development of COVID-19 vaccines and the widespread distribution and administration of vaccines have led to a decrease in the spread of the virus and a subsequent reduction in its impact. However, the rapid development of the COVID-19 vaccine, which would have taken at least a decade in the past (Lombard, Pastoret and Moulin, 2007), has raised concerns about its safety and long-term effects. These concerns are evident in social media, where there are numerous debates and instances of misinformation regarding the vaccine. This research project aims to employ sentiment analysis on Twitter comments related to COVID-19 vaccines in order to gain an understanding of the public's perspectives and opinions.

The objective of this study is to generate multiple advantages for the community, including:

- By gaining a comprehensive understanding of public opinion and attitudes towards COVID-19 vaccines, health authorities can modify their policies accordingly. This may include devoting additional resources to educate the public about the importance of vaccination.

- The results of this research can help to diminish misinformation about COVID-19 by providing scientifically sound data.
- A positive outcome from the sentiment analysis can contribute to an overall more positive perception of vaccination among the public.
- This study constitutes a contribution to the field of sentiment analysis and may generate new insights or raise further questions for future research on the subject.

The success of this research project is contingent upon several factors:

- The availability of a suitable dataset for analysis. Access to twitter data using Twitter's streaming API and Kafka is needed for real time data availability.
- Access to a streaming library like Apache Flink or Apache Spark Streaming to process the tweets and perform real-time analysis.
- Access to a cloud-based computing service such as AWS (Amazon Web Services) or GCP (Google Cloud Platform) to scale the infrastructure to avoid overload issues.
- To build a real-time data pipeline Python library Tweepy is needed.
- To process the data, sentiment analysis tools are necessary. Two such tools, the open-source Natural Language Toolkit (NLTK) package and Valence Aware Dictionary and Sentiment Reasoner (VADER), will be utilised in this data science project.

When conducting this research, certain assumptions will be made:

- Sentiment analysis is an appropriate technique for analysing the public attitude towards COVID-19 vaccines.
- Twitter data accurately reflects the public opinion, despite the possibility of certain individuals or groups using the platform for marketing or propaganda purposes, which may impact data quality.
- The proposed model is capable of detecting and processing multiple languages with a high degree of accuracy.
- The model exhibits a high level of accuracy in its classification of tweets into positive, negative, or neutral sentiments.

# 2. Literature Review

Sentiment analysis, also referred to as "opinion mining" or "emotion artificial intelligence," refers to the methodical identification, extrication, evaluation, and analysis of emotional states and subjective information using natural language processing (NLP), text mining, computational linguistics, and bio measurements (Alsaeedi and Khan, 2019). Sentiment research often examines the viewpoint expressed in client materials, such as online polls, reviews, and social media platforms. Sentiment analysis differentiates and categorises the author's viewpoint expressed in a given segment of text concerning the premise's foundational topic. The primary purpose of sentiment analysis is to establish the rate of polarity by which the author's tone in a corpus may be classified as positive, negative, or neutral (Praveen, Ittamalla and Deepak, 2021).

- Document-level sentiment classification: The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment (Turney, 2002; Pang and Lee, 2008, p. 2). This level of analysis assumes that each document expresses opinions on a single entity (e.g., a single product). Thus, it is not applicable to documents which evaluate or compare multiple entities.
- Sentence-level sentiment classification: The task at this level goes to the sentences and determines whether each sentence expresses a positive, negative, or neutral opinion.
- Entity and aspect (feature) level sentiment classification: Both the document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion). An opinion without its target being identified is of limited use. Realising the importance of opinion targets also helps us understand the sentiment analysis problem better. This level is also known as "perspective-level assessment grouping."

There is currently a vast amount of opinionated material available on the Web through social media. A lot of research would not have been possible without this information. Unsurprisingly, sentiment analysis's inception and rapid growth followed those of social media. In fact, sentiment analysis is currently at the core of social media research (Liu, 2012). The way individuals get information from people or organisations they are interested in has been shaped and transformed by TWITTER, a well-liked micro-blogging service throughout the world. Users of Twitter may

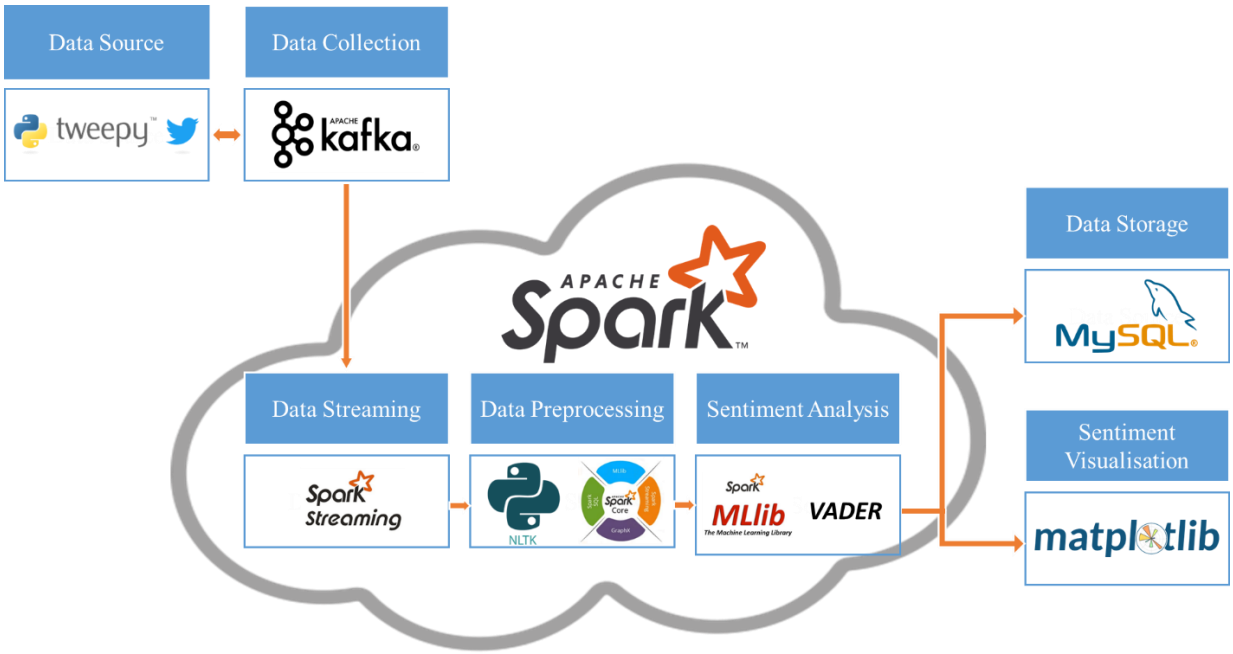
post tweets, or status updates, to let their followers know what they are thinking, doing, or what is going on in the world. By responding to or reposting another user's tweets, users may also communicate with one another. Twitter has grown to be one of the biggest social networking websites in the world since its founding in 2006 (Li, Dombrowski and Brady, 2018). Because to its numerous uses, mining users' stated attitude polarity in Twitter messages has emerged as a popular research topic in light of the ever-increasing volume of data accessible from Twitter (Tang *et al.*, 2015). For instance, various algorithms have been created to propose political election plans by evaluating the sentiment polarisation of Twitter users towards political parties and politicians (Wang *et al.*, 2012; Paul *et al.*, 2017). Also, businesses may quickly and efficiently track consumer opinion about their products and brands by employing Twitter sentiment analysis (Bravo-Marquez, Frank and Pfahringer, 2016).

A Twitter message's sentiment polarity can be classified as positive, neutral, or negative using sentiment analysis on the data from Twitter. Directly utilising conventional text sentiment analysis techniques is one way to undertake sentiment research on Twitter (Pang and Lee, 2006).

# 3. Architecture & Solution Design

The methodology for this project is illustrated in the below figure and further explained in this section.

*Figure 1: Architecture & solution design*




*Source: Compiled by the authors*

## 3.1. Data source

The first step in this project will be to collect tweets related to the topic of interest, which is different COVID-19 Vaccines. We will use the Tweepy API library, an open-source python package that allows us to access the Twitter API conveniently using its own classes and methods (Chaudhary and Niveditha, 2021). The Twitter API provides access to the live stream of tweets or to search tweets posted in the past.

## 3.2. Data collection

Apache Kafka receives data in text format and saves it in records. It utilises a storage layer that functions like a highly scalable message-queuing system that can receive, store, and transmit data. To manage data, Apache Kafka employs the use of clustering. Once received, data is passed on to Apache Spark, which processes the data to produce the intended results (D'silva *et al.*, 2017).

### 3.3. Data streaming

Apache Spark reads the data from Kafka using its native Apache Spark Streaming API. Each DataStream contains one or more text data records (D'silva *et al.*, 2017). In our case, these are tweets.

### 3.4. Data pre-processing

The collected tweets will be pre-processed to remove irrelevant or redundant data. This stage includes text cleaning, tokenisation, stemming, stop-word removal, and other techniques to ensure data quality.

Pre-processing steps will be implemented using Python's Natural Language Toolkit (NLTK) library (Abiola *et al.*, 2023). Pre-processed data using the NLTK will be fed into the VADER lexicon and rule-based sentiment analysis tool. The following processes will be performed at this stage:

- Punctuation Removal: This technique involves removing punctuation marks from text data to standardise it. It ensures that similar words such as "Vaccine" and "Vaccine!" are treated the same way. However, the list of punctuation to exclude must be carefully chosen depending on the use case.
- Remove Numbers: This involves eliminating numbers that are not relevant to the research.
- Remove Duplicates: While stopwords are typically eliminated based on language data, in cases where a domain-specific corpus exists, removing frequent terms that do not hold significant value is also necessary to prevent duplication and ensure data accuracy.
- Remove URLs: When preparing data for analysis, URLs are often removed to eliminate any bias or irrelevance they may introduce. For example, removing URLs from the data is expected when analysing tweets.
- Remove Whitespaces: This involves eliminating extra spaces in the text, which increases text size and adds no value to the data.
- Lowering the Text: The lower casing is a common technique used to standardise text. It ensures that words with different cases are treated the same. In addition, the lower casing helps to reduce duplication and ensure accurate counts in text featurisation techniques like frequency.

- Stop Words Removal: This involves eliminating frequently used words from the text that provide no value to the study. Such words are either meaningless or insignificant. While a list of stop words exists, it should be carefully selected based on the research.
- Lemmatisation: This technique stems words while keeping their sense intact. A predefined dictionary checks the word context as it gets smaller.
- Tokenisation involves breaking the text into individual tokens such as phrases, words, or characters. It is essential for many NLP procedures during pre-processing, where text is converted into word tokens.

Further, we will use the Spark ML API facilitates to perform the data pre-processing, which feeds into the Machine learning algorithms for the sentiment analysis available in Apache Spark (Ahmed, 2020).

### 3.5. Sentiment analysis

The sentiment analysis stage involves using VADER and Apache Spark's native sentiment analysis tool to classify tweets into positive, negative, or neutral sentiments. We believe the combination of a rule-based and machine-learning sentiment analysis tool will produce more accurate results.

- VADER is a lexicon and rule-based sentiment analysis tool that uses a human-curated sentiment dictionary and rule-based heuristics to determine the sentiment of a piece of text and is specifically suited for detecting sentiments on social media (Çilgin *et al.*, 2022). We will use the VADER library in Python for this analysis.

  The results of VADER analysis provide information about the sentiment polarity of a given word, indicating the likelihood of it being categorised as positive, negative, neutral, or compound. VADER calculates a compound score by searching the text for recognisable emotional features, adjusting their intensity and polarity using predefined rules, summing up the scores of all features detected in the text, and finally, normalising the total score within the range of -1 to 1.

- Apache Spark: We intend to use the available Naïve Bayes machine learning algorithms in the Spark machine learning library. These algorithm implementations are optimised to run in a distributed and parallel manner over a cluster of computing and storage resources (Ahmed, 2020).

### 3.6. Data storage and sentiment visualisation

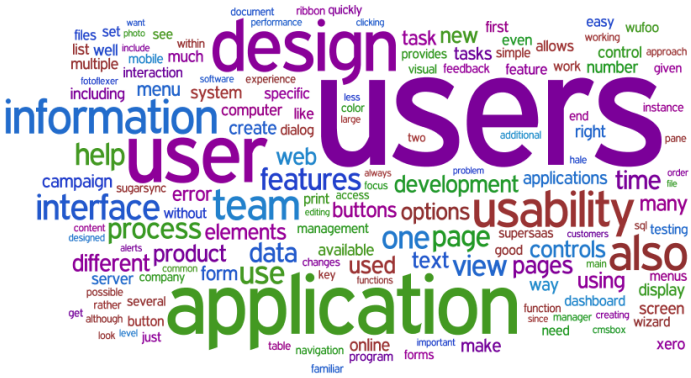
We will store sentiment scores in a Structured Query Language (SQL) database. For this, we intend to use the free, open-source SQL database called MySQL.

To visualise the sentiment analysis results, we will utilise the Matplotlib python library to generate graphs and charts. Matplotlib is a Python library for data visualisation.

# 4. Evaluation of sentiments

Our analysis will involve generating a Word Cloud, a tool used to explore the overall structure of texts, in this case, tweets. Word Cloud is a form of data visualisation that displays the words in a tweet or text on a chart, with more important words presented in larger fonts and less important words in smaller fonts or not shown at all. The image in Figure 2 below provides an example of a Word Cloud, which is not related to COVID-19 vaccines but is included here for illustrative purposes only.

*Figure 2: Sample Word Cloud*




*Source: Tag Cloud Examples* (Nielsen, 2009)

Further, we will assess the accuracy of our model. The accuracy score is a frequently used metric to evaluate the performance of a model. It measures the percentage of test data instances that the model correctly predicted. It is determined by dividing the number of correctly predicted instances by the total number of predicted instances (Saad *et al.*, 2021). The formula for calculating accuracy is:

$$Accuracy = \frac{T\,N + T\,N}{T\,N + T\,N + F\,P + F\,N}$$

# References


Abiola, O. *et al.* (2023) 'Sentiment analysis of COVID-19 tweets from selected hashtags in Nigeria using VADER and Text Blob analyser', *Journal of Electrical Systems and Information Technology*, 10(1), p. 5. Available at: https://doi.org/10.1186/s43067-023-00070-9.

Ahmed, M.A. (2020) 'Arabic Sentiment Analysis using Apache Spark', *International Research Journal of Innovations in Engineering and Technology*, 4(2), pp. 31–40.

Bravo-Marquez, F., Frank, E. and Pfahringer, B. (2016) 'Annotate-Sample-Average (ASA): A New Distant Supervision Approach for Twitter Sentiment Analysis', in *22nd European Conference on Artificial Intelligence (ECAI)*. *22nd European Conference on Artificial Intelligence (ECAI)*, IOS Press, pp. 498–506. Available at: https://doi.org/10.3233/978-1-61499-672-9-498.

Chaudhary, J. and Niveditha, S. (2021) 'Twitter Sentiment Analysis using Tweepy', *International Research Journal of Engineering and Technology*, 08(04), pp. 4512–4516.

Çilgin, C. *et al.* (2022) 'Twitter Sentiment Analysis During Covid-19 Outbreak with VADER', *AJIT-e: Academic Journal of Information Technology*, 13(49), pp. 72–89. Available at: https://doi.org/10.5824/ajite.2022.02.001.x.

D'silva, G.M. *et al.* (2017) 'Real-time processing of IoT events with historic data using Apache Kafka and Apache Spark with dashing framework', in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pp. 1804–1809. Available at: https://doi.org/10.1109/RTEICT.2017.8256910.

Li, H., Dombrowski, L. and Brady, E. (2018) *Working Toward Empowering a Community: How Immigrant-Focused Nonprofit Organizations Use Twitter During Political Conflicts*. ACM Publications. Available at: https://doi.org/10.1145/3148330.3148336.

Liu, B. (2012) *Sentiment Analysis and Opinion Mining*. Cham: Springer International Publishing (Synthesis Lectures on Human Language Technologies). Available at: https://doi.org/10.1007/978-3-031-02145-9.

Lombard, M., Pastoret, P.P. and Moulin, A.M. (2007) 'A brief history of vaccines and vaccination', *Revue Scientifique Et Technique (International Office of Epizootics)*, 26(1), pp. 29–48. Available at: https://doi.org/10.20506/rst.26.1.1724.

Nielsen, J. (2009) *Tag Cloud Examples, Nielsen Norman Group*. Available at: https://www.nngroup.com/articles/tag-cloud-examples/ (Accessed: 3 March 2023).

Pang, B. and Lee, L. (2006) 'Opinion Mining and Sentiment Analysis', *Foundations and Trends® in Information Retrieval*, 1(2), pp. 91–231. Available at: https://doi.org/10.1561/1500000001.

Pang, B. and Lee, L. (2008) 'Opinion mining and sentiment analysis', *Foundations and Trends in Information Retrieval*, 2(1–2), pp. 1–135.

Paul, D. *et al.* (2017) ‘Compass: Spatio Temporal Sentiment Analysis of US Election What Twitter Says!’, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery (KDD ’17), pp. 1585–1594. Available at: https://doi.org/10.1145/3097983.3098053.

Saad, E. *et al.* (2021) ‘Determining the Efficiency of Drugs Under Special Conditions From Users’ Reviews on Healthcare Web Forums’, *IEEE Access*, 9, pp. 85721–85737. Available at: https://doi.org/10.1109/ACCESS.2021.3088838.

Tang, D. *et al.* (2015) ‘A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification’, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11), pp. 1750–1761. Available at: https://doi.org/10.1109/TASLP.2015.2449071.

Turney, P.D. (2002) ‘Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews’. Available at: https://doi.org/10.48550/arXiv.cs/0212032.

Wang, H. *et al.* (2012) ‘A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle’, in *Proceedings of the ACL 2012 System Demonstrations*. Jeju Island, Korea: Association for Computational Linguistics, pp. 115–120. Available at: https://aclanthology.org/P12-3020 (Accessed: 21 February 2023).

*WHO Coronavirus (COVID-19) Dashboard* (2023). Available at: https://covid19.who.int (Accessed: 3 March 2023).