Keith Hammond-sba22421

Covid-19 Vaccine Tweets Sentiment Analysis and Time Series Forecast

Word count-3168

**Executive Summary:**  
The Covid 19 epidemic shook the very core of global societies, enforcing the need for effective vaccines for protection against the virus. Vaccine’s introduction sparked discussions across the world and notable voicing on twitter. Gaining understanding of the sentiment in regards the vaccine can provide rich information towards public attitudes for effective decision making. Sentiment analysis conducted through the paper gave valuable insight for what current available data has to offer but to gauge the future understanding time series forecasting had to be implemented as to identify potential patterns for predicting the future sentimental landscape in regards the vaccine. This research paper aimed to show public perception on Covid-19 vaccines and identify influential sentiment dynamics that can be used for public health campaigns for promoting vaccine acceptance during another potential epidemic The sentiment peaks coinciding with the rise of the Delta variant of covid reinforced the need of proactivity in implementing communication to address public concerns or misconceptions to mitigate negative sentiment during an ongoing pandemic.

**Introduction:**  
The global impact of the COVID-19 pandemic has been unprecedented, demanding the development and fast deployment of effective vaccinations to combat the virus. The introduction of COVID-19 vaccinations has provoked extensive discussion. Understanding public opinions and attitudes about these vaccinations is critical for policymakers, public health officials, and researchers to help in informed decision-making and successful communication tactics.

Sentiment analysis provides an investigating into measuring sentiment conveyed in text data. Sentiment analysis gives insights into public opinion on certain issues by analysing vast amounts of textual data. Analysis in the context of COVID-19 vaccinations can assist evaluate general attitude regarding the vaccines, identify dominant sentiments, and find shifts or patterns over time.  
Sentiment research alone just gives a picture of present attitudes; understanding how feelings develop over time is also important. Time series forecasting will estimate future values based on previous trends. Time series forecasting models may give significant forecasts on the future sentiment landscape by studying prior sentiment data and recognising trends, correlations, and seasonality, enabling preventative measures and targeted actions. The purpose of this research study is to investigate the use of sentiment analysis in assessing public attitude about COVID-19 vaccinations, and to expand the analysis by utilising time series forecasting techniques to project future sentiment trends. This study aims to contribute to the expanding body of information on public views of COVID-19 vaccinations by utilising sentiment analysis and time series forecasting, providing insights that can benefit in legislation, communication tactics, and public health activities. The project will include gathering and analysing enormous amounts of textual data from many sources, using cutting-edge sentiment analysis techniques, and putting time series forecasting models into action. This study's findings will offer insight on current sentiments concerning COVID-19 vaccinations, identify variables impacting sentiment dynamics, and estimate the future sentiment landscape.

The findings of this study have the potential to improve our understanding of public perception of COVID-19 vaccinations, allowing for evidence-based decision-making and proactive interventions. We may acquire significant insights into the people's developing ideas and attitudes by using the capabilities of sentiment analysis and time series forecasting, contributing to effective communication strategies and focused public health campaigns during these difficult times.

1- **Data Storage and processing activities carried out in the Spark Environment**  
Data storage and processing are critical components of performing efficient sentiment analysis and time series forecasting jobs. In this sample response, we will go over the complete process of preparing data and executing data processing operations in an Ubuntu environment using Apache Spark for twitter sentiment analysis and time series forecasting.

Data Gathering:  
The first stage in data preparation is gathering the necessary information. The Author tried using Twitter API acquisition of tweets for the assignment but due to the limitations of bestowed upon the application due to recent updates to API use by Elon Musk it would only generate a weeks’ worth of tweets so the Author researched a valuable dataset that met the criteria of a year of tweets minimum for those in relation of tweets based on the covid-19 vaccine search for the sentiment analysis due to the updated twitter system limitation of our initial API results. The dataset that fulfilled the criteria was discovered and included 384128 tweets. (Hiksuresh, 2023)  
Within the data it was researched to acquire a set that would specifically have historical data points within the data for the target variable to be able to perform a forecast time series.

Data Cleaning:  
The Unprocessed tweets from Twitter or other sources frequently contains noise, errors, or useless information. Cleaning the data is critical to ensuring the accuracy and quality of the subsequent analysis. The Author filtered out retweets, removed URLs or special characters (Emojis), and handled missing (null) data correctly in this stage.

Text Pre-processing:   
To analyse tweet sentiment, textual information inputs had been converted into numerical representations. Tokenization, lowercasing, stop word removal and managing emojis and hashtags were all text pre-processing undertaken at this stage. The pre-processed text data was then translated into an analysis-friendly format for Time Series Data Formatting: The Time series forecasting required organizing the data in a specific format due to Phantoms framework, The data should consist of a timestamp column and the corresponding target variable values. In the authors case the data was converted to including hours to just the year, month, date and target variable was the selected as the compound for future forecast of neutral, positive and negative sentiments.

Training and Evaluation:  
The pre-processed data was then divided into multiple datasets for analysis. The analysis of the data was used for assessment in the Spark environment. For time series forecasting, the external Prophet library was utilised to develop forecasting models that provided the author with the results needed for one week. one month and three months. Making use of this approach the data was able to predict the compound of the sentiment of the text.

Scaling and Distributed Computing:  
Spark's distributed computing features enable the processing of large-scale datasets by distributing the data among numerous nodes in a cluster. Sentiment analysis and time series forecasting operations may be scaled to accommodate huge data scenarios easily by exploiting Spark's parallel processing and distributed techniques. (Delteil,2023)   
The storage and processing of data is critical to the performance of sentiment analysis and time series forecasting jobs. We looked at the detailed process of data preparation in this sample response, including data collection, cleaning, text pre- processing, and time series formatting. Then the Author spoke about how to use Apache Spark in an Ubuntu environment to accomplish data processing activities such importing data into Spark environment, data transformation, feature engineering, algorithm configuration and assessment. These operations were expanded to handle large-scale datasets successfully by utilizing Spark's distributed computing capabilities, enabling accurate sentiment analysis and dependable time series forecasting. (Albaldawi and Almuttairi, 2020)

2-**Rational and Justification**   
Throughout this paper the Author has been asked to provide the rational and justify the choices in regards to data processing, storage on data, programming language of choice, machine learning models and algorithms used to present the findings of the work conducted, the following sections detail as such in the considerations the Author took for analysing the twitter data on COVID-19 vaccines for sentiment analysis and time series forecasting:

* Data Processing:

For the Author to efficiently analyse a huge volume of Twitter data, a distributed computing system capable of parallel processing over a multitude of devices is required. For this project, the Apache Spark framework was the selected processing. PySpark’s high-level distributed data processing works well with Python. It dramatically accelerates data processing procedures by leveraging Spark's in-memory computing capabilities. Furthermore, Spark is fault-tolerant and scalable, making it ideal for handling big data tasks such as transformation, feature engineering, model training, and assessment. These operations make it ideal for expanding to handle large-scale datasets successfully by utilizing Spark's distributed computing capabilities, enabling accurate sentiment analysis and dependable time series forecasting as being asked to perform for the report. The ease of usage and connection with Python of PySpark were key advantages. Python being such a popular programming language in the data science field because of its rich libraries and frameworks allows the Author use PySpark to achieve the power of Spark's distributed processing while taking use of Python's extensive ecosystem for data manipulation, analysis, and visualisation.

* Storage:

Choosing the appropriate data storage solution is critical considering the large volume of Twitter data involved in the project, 384128 tweets. Distributed and scalable storage systems were the preferred choice for efficient handling of big data.

HDFS, as a distributed file system, enabled the storing and processing of the massive datasets over a multitude of machines. It has high throughput and fault tolerance, making it a good choice for huge data storage. HDFS's scalability and parallel processing capabilities are ideal for this project's needs, allowing for efficient data storage and retrieval. The decision of using HDFS had been influenced by aspects such as data structure, query patterns, and the requirement for real-time analytics. HDFS was preferable as the emphasis was on analysing old data using complicated queries.

* Programming Language Choice:

The Author selected the use of Python as being chosen for the programming language of this project due to its popularity, large library, and active community that are constantly updating and adding to the libraries freely available to the public making it highly beneficial for its versatility and simplicity of use, Python is the language of choice for data analysis, machine learning, and natural language processing (NLP). The combination of Pyspark with Python enabled the use to integrate seamlessly with Sparks distribution capabilities, making it the authors preferred choice for the task bestowed upon thee. A multitude of available libraries improve the processing and analysis of the dataset. Python was the ideal choice on the basis of its ability to produces fast processing , data visualisation and machine learning outcomes The inclusion on NLP libraries for python provides the needed tools to produce pre-processing of the dataset text , create sentiment analysis and forecasting of the covid-19 vaccine tweets gathered. (Paul, 2021)

* Machine Learning Models and Algorithms:

Prophet is a machine learning algorithm developed by Facebook's Core Data Science team. It is specifically designed for time series forecasting tasks. While Prophet incorporates machine learning techniques, it is not a traditional machine learning algorithm in the sense of classification or regression. Prophet is based on an additive model that decomposes time series data into trend, seasonality, and holiday components. It uses a Bayesian framework and employs a piecewise linear or logistic growth curve for modelling the trend. The seasonality is modelled using Fourier series expansions, and holidays are incorporated as additional regressors.

Prophet is known for its simplicity and ease of use, making it accessible to users with less experience in time series forecasting. It automates many of the steps involved in building a forecasting model, such as handling missing data, outlier detection, and trend detection.

By providing a straightforward interface and customizable parameters, Prophet allows users to quickly generate accurate forecasts for a variety of time series data, including business metrics, sales, stock prices, and more. The selection of using prophet for time series forecasting model was influenced by variables such as data properties, the existence of seasonality or trend, and the required prediction accuracy. The choice between Prophet and other machine learning algorithms like SVM, logistic regression, naive Bayes, and random forest for sentiment analysis of tweets was done so on the premise of itsability of categorizing time series of tweets**.** Prophet is particularly competent for the application of time series forecasting tasks, when the importance of the task is on the prediction of future values based on previous values. When incorporating the primary aim to analyse sentiment patterns in tweets over time and forecast future events.

However, if the primary goal is to classify individual tweets as positive, negative, or neutral without a strong emphasis on time series analysis, the other algorithms mentioned (SVM, logistic regression, naive Bayes, and random forest) are more commonly used for general sentiment analysis tasks. These algorithms have been frequently used in text classification and sentiment analysis settings, and they may perform well when trained on a big and varied labelled dataset. (Kutzkov, 2023)

* Environment:

The Ubuntu operating system provided a robust and extensively utilised platform for data processing and analysis. Because of its communication with many data processing frameworks and libraries, including PySpark and Python itself, Ubuntu was the popular choice for the Author. Ubuntu's stability and vast community support offer a dependable environment for project execution.   
Then from within Ubuntu the Spark environment was chosen as Spark contains a scalable machine learning library (MLlib) which covers a multitude of algorithms for natural language processing (NLP) workloads, including that of the assignment’s main focus, sentiment analysis.  
The distributed implementation of MLlib enables rapid training and deployment of sentiment analysis models on huge datasets. It also simplifies the development process by supporting feature extraction, model assessment, and hyperparameter tweaking.

Overall, the choices made in this project for data processing, storage, programming language, machine learning models, and algorithms were motivated by the need to efficiently handle large volumes of data, achieve accurate sentiment analysis, and provide reliable time series forecasting for COVID-19 vaccine tweets over the time span of the collected data. PySpark, Python, HDFS, NLP packages, and machine learning algorithms have proved useful in the construction of the tasks of creating sentimental and time series forecasts which provided a solid framework for extracting insights from the data.

3.**Comparative Analysis of 2 Databases using YCSB**

The Author used the Yahoo! Cloud Serving Benchmark (YCSB) to do a quantitative examination and comparison of two database systems, MySQL and MongoDB. This investigation sought to assess the performance and applicability of both systems. The quantitative analysis compares the two databases with a record count increased from 1,000 to 10,000. The workload distribution was configured to be 80% reads and 20% writes. The performance metrics produced the following results: MongoDB's throughput was 3589 operations per second, while MYSQL's was 404. MongoDB's better throughput was due to its document-oriented design and indexing capabilities.  
MongoDB was roughly ten times faster on those criteria in terms of latency. MongoDB's result is 328, whereas my SQL result is 3021. Lower latencies were made possible by MongoDB's optimised data storage format and indexing methods.

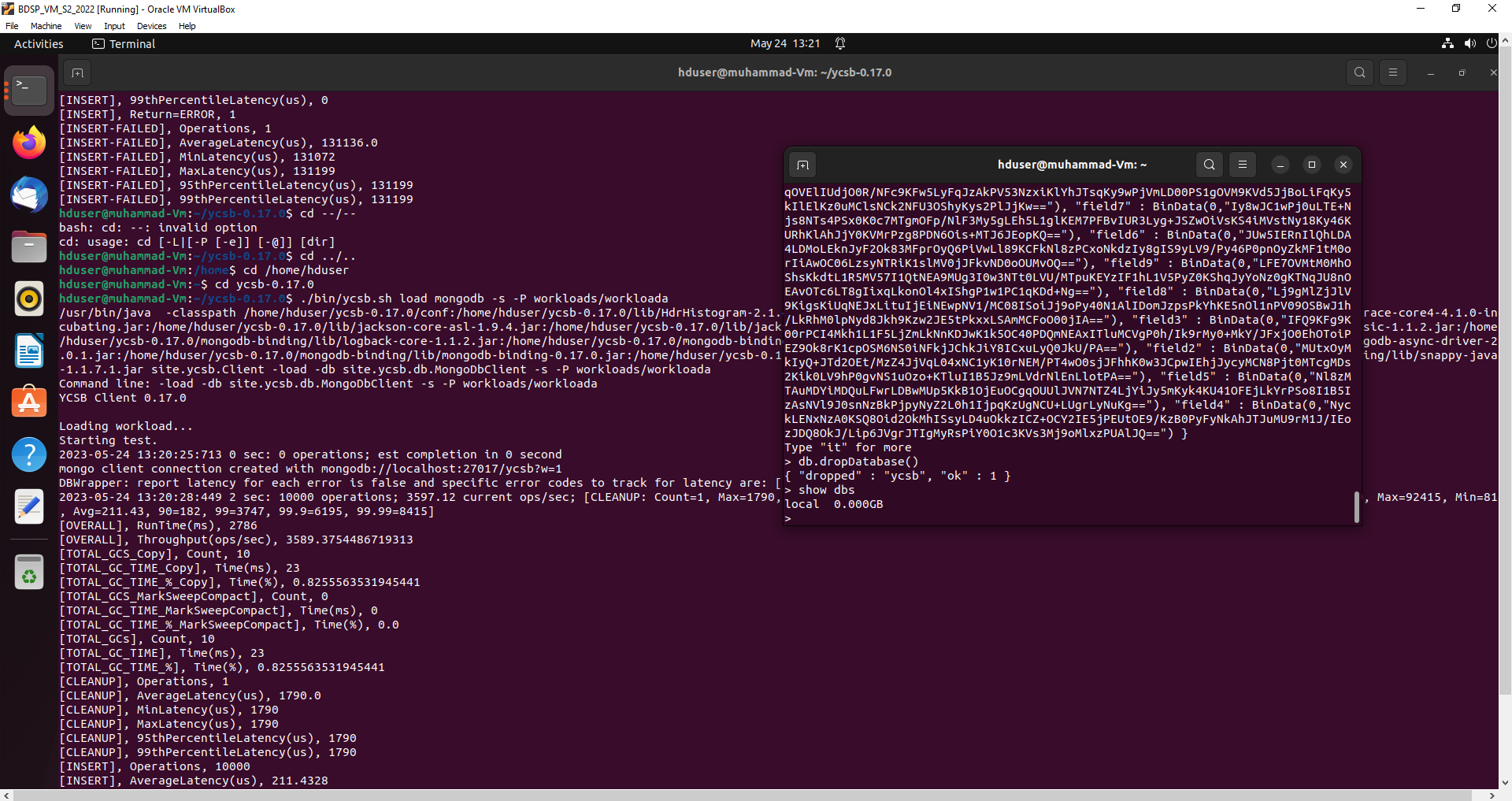
Because of its faster throughput and reduced latency, MongoDB is recommended for read-intensive applications. Both databases handled the workload combination of 80% reads and 20% writes effectively. However, the runtime in mongos is substantially better at 2,786ms compared to MYSQL’s findings of 24,711ms. The quantitative comparison using YCSB offered useful insights into the performance and appropriateness of MYSQL and MongoDB for our unique use case. The test results are available in the figures below:  
  
  
Figure 1. MongoDB results and Connection terminals inside VM  
  


Figure 2. MySQL results within VM terminal

4- **Analysis of Changing Sentiments over Time in COVID-19 vaccine**

The COVID-19 vaccine tweets were examined through sentiment labels assigned to the dataset items to see how sentiment has changed over time. Dataset pre-processing took place with the COVID-19 vaccine dataset includes textual information as well historical dates. The Author cleaned the text entries and removed any irrelevant columns before pre-processing the dataset (e.g., deleting punctuation, special characters, and lowercasing).

For Sentiment Analysis the Author used a pre-trained sentiment analysis model to analyse the sentiment. This approach tags each text item with a sentiment label, categorising it as positive, negative, or neutral. The sentiment analysis model was trained on the compound data, it was then used in a combination of natural language processing techniques and a machine learning algorithm. This allowed us to estimate sentiment labels for the dataset's entries.  
After assigning sentiment labels to each tweet the author then examined the sentiment distribution across time by separating the dataset into time periods and estimated the score of positive, negative, and neutral feelings through compound analysis. Through this process it was found that though the overall consensus of the collective of sentiment was positive within the time frame in regards the vaccine, it was at its most prevalent when in conjunction with the timing of the vaccine release and the roll out of the vaccine , this can be seen within the graphs presented below:

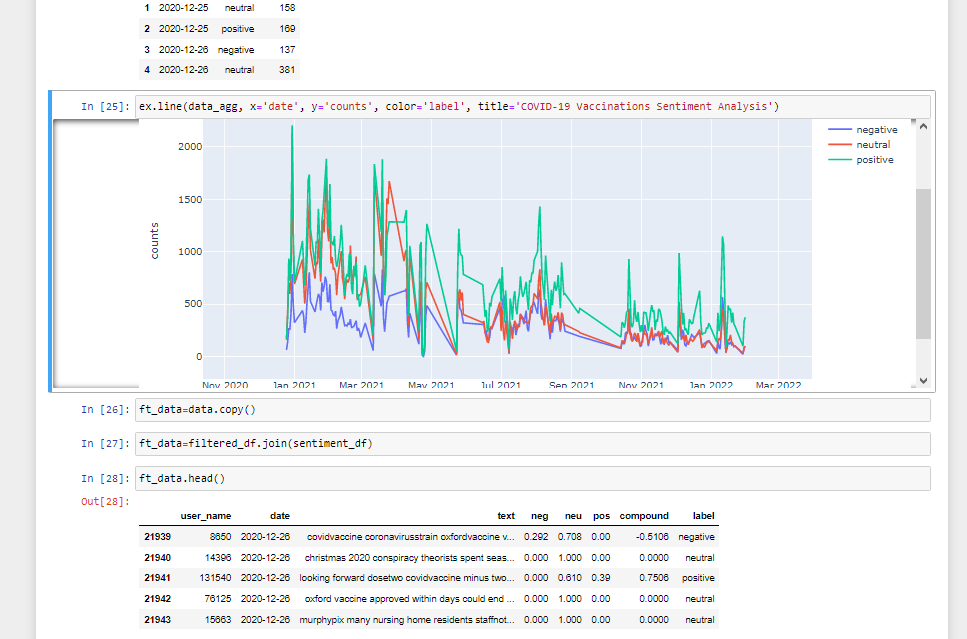
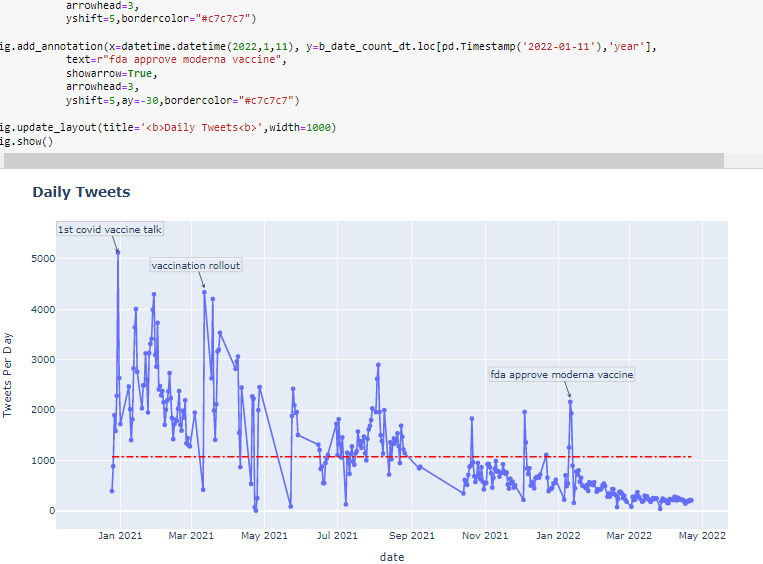
  
Figure 3- Tweet sentiment over the time period  
  
  


Figure 4- Daily tweet activity and cause for the activity

**5- Forecast of the Sentiment at 1 week, 1 month and 3 months Going Forward**

When it comes to time series forecasting, choosing the appropriate model was crucial for accurate predictions. While there is an array of models available, Prophet stood out as the powerful tool for its flexibility and capability to handle various time series patterns. To show justification of the Authors choice of time series he has highlighted a few key advantages over another popular choice for time series forecasting- Autoregressive Integrated Moving Average (ARIMA). Prophet offers multiple advantages for time series forecasting over ARIMA. (Tchoudi, 2022)  
Its ability to handle seasonality, flexibility in trend modelling, consideration of multiple components, robustness against missing data and outliers, interpretability, scalability, and community support made it an attractive choice for forecasting sentiment based for the author on a year's worth of tweets.

A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer screen

Description automatically generated with medium confidence  
Figure 5- Forecast trends Figure 6- Scatterplot depicting forecast trends

The results of the forecasts showcased a positive trend moving forward with sentimental forecasting going forward. The same trend can be seen in the results of the graphs within week 1 , 1 month and 3 months, the confidence interval expands the longer the forecasting goes into the future as the uncertainty regarding the sentimental values moves forward.

6- **Conclusion**  
Throughout the study of the COVID-19 vaccine twitter dataset revealed changes in sentiment over time. The initial negative mood turned to good sentiment, implying that individuals gradually adapted developed resiliency to the hesitation of getting the vaccine. However, the mood remained attentive to continuing epidemic events and conditions hence the neutrality in most of the sentiment in the time period of the dataset. Understanding the evolving sentiment during the epidemic can give policymakers, public health professionals, and researchers with useful insights. It can identify places where public messages and actions may be required to address issues and sustain public trust by analysing sentiment patterns if such a pandemic develops again in the future. The dataset used for this analysis represents a single source, and sentiment may vary across different platforms or demographics. Expanding the analysis to include multiple datasets or sources could provide a more comprehensive understanding of sentiment dynamics. Additional contextual information, such as location or demographics, could enhance the analysis by uncovering regional or group-specific sentiment patterns.

Analysing the sentiment changes over time within the COVID-19 vaccine dataset revealed the evolving attitudes and perceptions surrounding the covid vaccines. The research results provide valuable insight into public attitude at various points of the epidemic, enabling for a better understanding of the publics opinions on the events that transpired.

* During the early phases of the epidemic, public opinion appeared to be overwhelmingly vocal. People voiced the need for a cure.
* As time passed and more people learned about the vaccine, public opinion began to alter. As people and communities adapted to the new normal, positive attitudes such as the sentiments remained positive throughout the dataset lifespan, influenced by events such as the emergence of new vaccines, vaccination programmes, and government regulations.
* Time series forecast a significant increase in sentimental attitude as reported during the 1 week 1 month and 3 months forecast. These periods suggest as the knowledge and where with all of vaccinations become more commonly known that the attitude towards them steadily increases.

**References**

AI Perspective (2022) ‘*12 twitter sentiment analysis algorithms compared’*, *AI Perspectives*. Available at: https://www.aiperspectives.com/twitter-sentiment-analysis/   
Accessed : 16/5/2023

Albaldawi, W. and Almuttairi, R. (2020) ‘*Near real time twitter sentiment analysis and visualization’*. Available at: https://iopscience.iop.org/article/10.1088/1757-899X/928/3/032044   
Accessed: 18/5/2023

Arora, S. (2023) ‘*Sentiment analysis using Python*, *Analytics Vidhya’*. Available at: https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/   
Accessed: 18/5/2023

Benchant (2021) ‘*The Ultimate YCSB Benchmark Guide’*, *Benchant.com*. Available at: https://benchant.com/blog/ycsb Accessed on: 18/5/2023

Delteil, C. (2023) ‘*Large-scale sentiment analysis with pyspark’*, *Medium*. Available at: https://pub.towardsai.net/large-scale-sentiment-analysis-with-pyspark-bdccf9256e35   
Accessed: 16/5/2023

Elzayady, H., Badran, K. and Salama, G. (2018) ‘*Sentiment analysis on Twitter data using Apache Spark Framework’*. Available at: https://www.researchgate.net/publication/331106576\_Sentiment\_Analysis\_on\_Twitter\_Data\_using\_Apache\_Spark\_Framework   
Accessed: 14/5/2023

Furqan, M. (2018) ‘*Big data - twitter’s tweets sentiment analysis using Apache Spark Streaming’*, *LinkedIn*. Available at: https://www.linkedin.com/pulse/big-data-spark-streaming-twitter-sentiment-analysis-muhammad-furqan   
Accessed: 16/5/2023

Goyal, G. (2023) ‘*Twitter sentiment analysis using Python: Introduction & techniques*, *Analytics Vidhya’*. Available at: https://www.analyticsvidhya.com/blog/2021/06/twitter-sentiment-analysis-a-nlp-use-case-for-beginners/   
Accessed: 18/5/2023

Hayashi, H. (2017) ‘*Is prophet really better than Arima for forecasting time series data?’*, *Medium*. Available at: https://blog.exploratory.io/is-prophet-better-than-arima-for-forecasting-time-series-fa9ae08a5851   
Accessed: 17/20/2023

Hiksuresh, Kaus (2023) ‘*Covid vaccine Eda’*, *Kaggle*. Available at: https://www.kaggle.com/code/kaushiksuresh147/covid-vaccine-eda/input   
Accessed: 14 May 2023.

Hoekstra, D. (2021) ‘*How to build a Twitter sentiment analysis tool’*, *freeCodeCamp.org*. Available at: https://www.freecodecamp.org/news/how-to-build-a-twitter-sentiment-analysis-tool/   
Accessed: 16/5/2023

Jordan, M. (2022) ‘*YCSB’*, *ScyllaDB*. Available at: https://www.scylladb.com/glossary/ycsb/   
Accessed: 20/5/2023

Kutzkov, K. (2023) ‘*Arima vs Prophet vs LSTM for time series prediction’*, *neptune.ai*. Available at: https://neptune.ai/blog/arima-vs-prophet-vs-lstm   
Accessed: 18/5/2023

Mhatre, S. (2022) ‘*Sentiment analysis with python’*, *Simple Talk*. Available at: https://www.red-gate.com/simple-talk/development/data-science-development/sentiment-analysis-python/ Accessed: 20/5/2023

Muluberhan-Berhe, N. (2020) ‘*Congressional tweets: Using sentiment analysis to cluster members of Congress in PySpark’*, *Medium*. Available at: https://medium.com/analytics-vidhya/congressional-tweets-using-sentiment-analysis-to-cluster-members-of-congress-in-pyspark-10afa4d1556e Accessed: 15/5/2023

Pascual, F. (2022) ‘*Getting started with sentiment analysis on Twitter’*, *Hugging Face – The AI community building the future.* Available at: https://huggingface.co/blog/sentiment-analysis-twitter Accessed: 20/5/2023

Paul, S. (2021) ‘*Python sentiment analysis tutorial’*, *DataCamp*. Available at: https://www.datacamp.com/tutorial/simplifying-sentiment-analysis-python   
Accessed: 14/5/2023

Stamatelou, E. (2020) ‘*Sentiment analysis on streaming Twitter data using SPARK Structured Streaming & Python’*, *Medium*. Available at: https://towardsdatascience.com/sentiment-analysis-on-streaming-twitter-data-using-spark-structured-streaming-python-fc873684bfe3   
Accessed: 15/5/2023

Tchoudi, W. (2022) ‘*Forecasting valuation comparing Arima vs prophet’*, *LIGS University*. Available at: https://ligsuniversity.com/blog/forecasting-valuation-comparing-arima-vs-prophet#:~:text=The%20results%20of%20the%20comparison,a%2011%25%20error%20in%20prediction. Accessed: 18/5/2023

Xiaoyuan, W. (2021) ‘*Topic modeling and sentiment analysis on Twitter data using Spark’*, *Medium*. Available at: https://towardsdatascience.com/topic-modeling-and-sentiment-analysis-on-twitter-data-using-spark-a145bfcc433   
Accessed: 15/5/2023

**Github Repository Link** <https://github.com/sba22421/Ca2combouploadGIT/blob/main/Ca2SenitmentForecastUpload.ipynb>