# Covid-19 Vaccine Tweets Sentiment Analysis Using Apache Spark and Mongo DB

Hlaing Hlaing Win
Information Technology Supporting and Maintenance Department
*University of Information Technology,Yangon,*
*hlainghlaingwin@uit.edu.mm*

Tun Ye Min
*University of Information Technology,Yangon,*
*tunyeminn@uit.edu.mm*

Kyawt Kyawt San
Faculty of Information Science
*University of Information Technology,Yangon,*
*kyawtkyawtsan@uit.edu.mm*

Khin Nyo Nyo Theint
*University of Information Technology,Yangon,*
*khinnyonyotheint@uit.edu.mm*

Aung Khant Myat
University of Information Technology,Yangon,
aungkhantmyat@uit.edu.mm

Khin Ei Ei Chaw
Faculty of Information Science
*University of Information Technology,Yangon,*
*eichaw@uit.edu.mm*

***Abstract*— The development and distribution of vaccines have become pivotal milestones in our collective journey toward normalcy in an era defined by the COVID-19 pandemic. Tweets become an important real-time hub for public discourse as COVID-19 vaccine discussions unfold across different platforms. With Natural Language Processing (NLP) and data analysis, our project explores this vast and dynamic landscape of Twitter for insights and sentiment patterns related to COVID-19 vaccinations. Sentiment analysis, often referred to as opinion mining, is a technique within natural language processing (NLP) that focuses on deciphering the emotional tone or sentiment expressed in textual data. It involves analyzing text and classifying it as either positive, negative, neutral, or somewhere along a spectrum of emotions. The relationship between sentiment and external factors is uncovered by using Correlation Analysis . A further investigation into user behavior is also conducted to identify how different categories of user's express sentiment about COVID-19 vaccines. As we navigate the intricacies of text data, Text Decomposition Analysis (TDA) is employed to reduce the dimensionality of the data and identify the key words that determine sentiment variance. As a final step, we perform a comprehensive sentiment analysis, calculating positive, negative, and neutral sentiments, along with polarity scores, and assessing its accuracy using machine learning.**

**Keywords—sentiment analysis, machine learning, vaccine tweets**

## I. Introduction (*Heading 1*)

Since the COVID-19 pandemic was declared in March 2020, there has been a concerted international effort to create and test COVID-19 vaccinations. Although COVID-19 preventative methods have shown to be moderately helpful in restricting its spread, the outcome of the pandemic will depend greatly on the protective and durable immunity provided by vaccination. To achieve any level of herd immunity, it is predicted that at least 70% of the population will need to be immunized (Orenstein and Ahmed, 2017; Aguas et al., 2020). Public support for vaccination is crucial to achieving this. Therefore, it is crucial to comprehend the public's attitudes on vaccination and, consequently, their willingness to receive vaccinations.

In order to increase vaccination rates among the public, it may be helpful to address vaccine opposition and develop vaccine trust (Ferrer and Ellis, 2019). This can be done by creating vaccine-promotional messages that are tailored using information learned about vaccine attitudes and opinions. According to community characteristics including demographics, income, and family or religious status, opinions on the COVID-19 immunization can vary (Lyu et al., 2020).

This paper mainly proposes and presents on Twitter sentiment's rich tapestry and reveal valuable insights into public sentiment dynamics and COVID-19 vaccines' impact on digital discourse. Through analysis of sentiment expressed on Twitter, our proposed system aims to help us understand how the general public views and feels about COVID-19 vaccines. . Furthermore, the proposed system can contribute to academic research on sentiment analysis, text mining, and the dynamics of public sentiment in response to health crises.

## II. Materials and Methods

Sentiment analysis can be approached using various methods, including: rule-based approach, machine learning-based approach, and hybrid approach. For this analysis, we have chosen the machine learning-based approach.

### *A. Nature of Tweets Data*

In our sentiment analysis of Twitter data, we encountered a variety of sentiments:

**Negative:** "Five years ahead of schedule? It's a pandemic! There was no pre-determined schedule." This tweet expresses a negative sentiment, highlighting frustration and criticism regarding an event not going as planned during the pandemic.

**Neutral:** "Yes, here is the link to the patient education information from #PfizerBioNTec." This tweet conveys a neutral sentiment, merely providing a link to patient education information without expressing any strong emotions.

**Positive:** "Happy and relieved to have the Pfizer BioNTech Covid vaccine. #GetVaccinated!" This tweet radiates positivity and relief, with the user expressing happiness and encouraging others to get vaccinated, reflecting a positive sentiment.

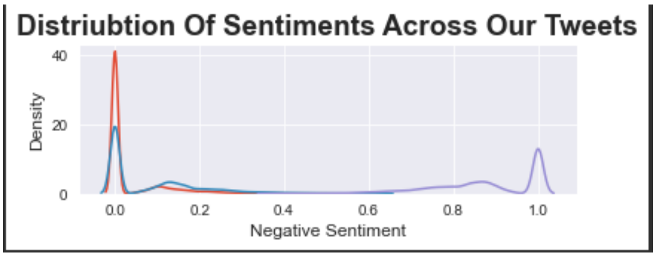

Figure 1. Tweets Data Distribution

According to Figure1, it is observed that the distributions of the sentiments follow a normal distribution; the negative and positive sentiments are very similar, proposing that there may be no significant differences in the strength of our data's positive and negative sentiments.

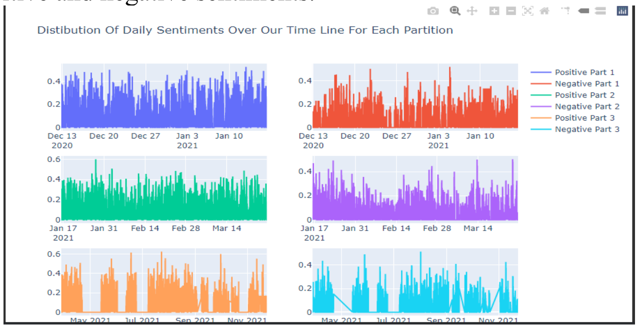

Figure 2. Tweets Data on Time based Analysis

Figure 2 shows that the tweets' sentiments do not meet stationarity requirements as to non-constant mean and variance.

## B. Data Set

Dataset is about the Covid19 Vaccines mainly about the Pfizer/Biotech Vaccine Tweets from Twitter API. This dataset is from Kaggle.com.

The dataset includes 11020 rows and 17 columns. The data type includes a mixed of textual data, categorical data, and numerical data. The dataset is in json format.

The features of the dataset are:

- id : Twitter user's id
- user_name : Twitter user name
- user_location : Location of the twitter user posting the tweet
- user_description : Twitter user's Twitter Profile Information
- user_created : The date user create twitter account
- user_follower : Number of followers of the user
- user_friends : Number of friends the user have on Twitter
- user_favourites : Number of tweets the user favorited
- user_verified : Whether the user is verified profile or not
- date : The date where the tweet is posted
- text : Text on the tweet
- hashtags : Hashtag on the tweet
- source : Device the tweet is posted
- Retweets : Number of retweets
- Favourites : Number of time the tweet had been favorited
- Is_retweet : Whether the tweet is the retweet post of others

## C. Data Storage

The dataset used for analysis is stored in MongoDB Atlas. Thus, the information is extracted from the database. MongoDB Atlas is used as the storage. By creating a free tier shared cluster (Cluster0), using AWS as cloud provider and choosing N.Virginia(us-east-1) region. The connection string is ***"mongodb+srv://khinezinthant:(password)@cluster0.fkq2fra.mongodb.net/"***. Thus, the dataset can be retrieved from teammates from different geographical locations to do separate analysis tasks.

# III. System Architecture for Vaccine Tweets Sentiment Analysis

The system architecture for the Vaccine Tweets Sentiment Analysis is shown in Figure 3 which includes data collection to gather relevant tweets, secure storage in MongoDB in JSON format, data preprocessing with PySpark for cleaning and transformation, and sentiment analysis using Text Blob and Logistic Regression. Matplotlib and Plotly are employed for data visualization, facilitating the creation of insightful visual representations. This architecture aims to provide valuable insights from tweet data while ensuring efficient data management and analysis.and not as an independent document.

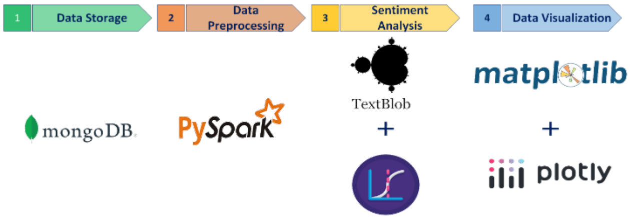

Figure.3. The System Architecture for Vaccine Tweets Sentiment Analysis

System Components of the Architecture are:

1. **Data Storage**
   - Purpose: Store tweet data in JSON format.
   - Technologies Used: MongoDB Atlas.
2. **Data Preprocessing**
   - Purpose: Clean and transform raw data.
   - Technologies Used: PySpark.
3. **Sentiment Analysis**
   - Purpose: Perform sentiment analysis on tweets.
   - Technologies Used: Text Blob, Logistic Regression.
4. **Data Visualization**
   - Purpose: Create visual representations of data.
   - Technologies Used: Matplotlib, Plotly.

## A. Data Pre-processing

By using pyspark as a data cleaning tool, the following tasks are performed: 1. Lower-casing, 2. Removing noise from tweet text using Regex, 3. Tokenization, 4.Stop word removal, 5.Duplicate removal, 6.Stemming, 7.Lemmatization.

## B. Accessing Sentiment Using Polarity Function of TextBlob

TextBlob is a Python library for natural language processing that includes a sentiment analysis feature. Sentiment polarity in TextBlob refers to the sentiment or emotional tone conveyed in a piece of text, and it is usually classified as either positive, negative, or neutral.

**Polarity Score:** Polarity is a critical aspect of sentiment analysis. The polarity score is a single numerical value that indicates the sentiment polarity of the text. It is measured using the 'sentiment. polarity' attribute from TextBlob, which provides a polarity score ranging from -1 (very negative) to 1 (very positive), with 0 indicating a neutral sentiment.

### C. Feature Extraction

For vectorization, Count Vectorizer is used to convert text data into numerical features that machine learning models can understand. It works by transforming a collection of text documents into a matrix of token counts, where each row represents a document, and each column represents a unique word or token in the entire corpus of documents. The values in the matrix indicate how many times each word occurs in each document.

In our implementation, ngram_range (1,2) is assigned in order to consider both unigrams (individual words) and bigrams (pairs of adjacent words) when extracting features from the text data. As a result, the total of 78583 features is extracted from our data.

### D. Logistic Regression

Logistic Regression is a supervised machine learning algorithm.It models the relationship between a dependent variable and one or more independent features by estimating the probability that an input belongs to a particular class.

### E. Support Vector Machine (SVM)

Another machine learning algorithm, Linear Support Vector Classifier (LinearSVC) is used for sentiment analysis. LinearSVC is a type of Support Vector Machine (SVM) that works well for text classification tasks, including sentiment analysis. It tries to find a hyperplane that best separates the data into different sentiment classes. The LinearSVC model is trained on the training data. During training, the model learns the relationships between the extracted features and the corresponding sentiment labels. After training, the model is tested on the testing set to assess its accuracy and performance by confusion matrix.

### F. Fine Tuning Parameters

Grid search is a fundamental technique for hyperparameter tuning in machine learning models like Support Vector Classifier (SVC) and Logistic Regression (LogReg) used in sentiment analysis. Hyperparameters, such as C (regularization strength), kernel type, degree (for polynomial kernels), and gamma (kernel coefficient), profoundly influence model performance. Grid search systematically explores predefined sets of hyperparameter values, like [0.01, 0.1, 1, 10] for C and ["linear", "poly", "rbf", "sigmoid"] for the kernel, training and evaluating models through cross-validation. The aim is to pinpoint the hyperparameter combination that delivers the best performance, often assessed by metrics like accuracy or F1-score. Once identified, these optimal hyperparameters are applied to train final models, ensuring they are finely tuned for the sentiment analysis task. Grid search simplifies the intricate process of optimizing hyperparameters, serving as an indispensable tool for enhancing model accuracy.

Grid search automates the process of hyperparameter tuning and helps find the hyperparameters that optimize the model's performance, ensuring that the model generalizes well to new data. It prevents the need for manual tuning, which can be time-consuming and less systematic.

## IV. Evaluation Metrics

Evaluation metrics are crucial for assessing the performance of sentiment analysis models. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These values are used to calculate various performance metrics, including precision, recall, and the F1-score as follows:

**Accuracy**: (TP + TN) / (TP + TN + FP + FN)

**Precision**: TP / (TP + FP)

**Recall (Sensitivity or True Positive Rate)**: TP / (TP + FN)

**Specificity (True Negative Rate)**: TN / (TN + FP)

**F1-Score**: 2 * (Precision * Recall) / (Precision + Recall)

According to the experiment of classification with hyperparameter tuning, the results of the two machine learning models, logistic regression and support vector machine are comparable and but logistic regression is superior with the accuracy of 85%.

Table1. Classification Report

| | precision | recall | F1-score | support |
|---|---|---|---|---|
| Negative | 0.86 | 0.32 | 0.46 | 226 |
| Neutral | 0.78 | 0.99 | 0.88 | 1021 |
| Positive | 0.94 | 0.82 | 0.87 | 862 |
| accuracy | | | 0.85 | 2109 |
| Macro avg | 0.86 | 0.71 | 0.74 | 2109 |
| Weighted avg | 0.86 | 0.85 | 0.83 | 2109 |

## V.Conclusion

Analyzing temporal sentiment trends over time can help identify when and how public sentiment has evolved in response to vaccine-related events, news, or policy changes. This can aid in trend prediction and preparation for future developments. Location-Specific Insights help pinpoint regions where sentiments are particularly strong, be it positive or negative. This information can guide targeted interventions or communication strategies in specific areas. User Behavior Analysis involve categorizing users by popularity and analyzing their sentiments can help in identifying key opinion leaders and understanding how different user groups engage with vaccine-related discussions. This knowledge can be useful for social media strategies. Policymakers can use sentiment analysis to gauge public reception of vaccination policies and make data-driven decisions to improve public health outcomes. In our analysis of Pfizer Twitter data, sentiment analysis proves invaluable for deciphering the emotional undercurrents within user-generated content.

## References

[1] Abd Rahim N, Rafie SM. Sentiment analysis of social media data in vaccination. Int J
2020];8(9).

[2] 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
ealthc 2020a];1–8

[3] Dataset: Pfizer Vaccine Tweets https://www.kaggle.com/datasets/gpreda/pfizer-vaccine-tweets

[4] Ferrer RA, Ellis EM. Moving beyond categorization to understand affective influences on real world health decisions. Soc Pers Psychol Compass 2019];13(11):e12502.

[5] Orenstein WA, Ahmed R. Simply put: vaccination saves lives. National Acad Sciences; 2017.