Sentiment Analysis on COVID-19 Vaccine Tweets

Amish Raj

Department of Computer Science

George Washington University

amish.raj@gwu.edu

Boomika Karuppaiah

Department of Computer Science

George Washington University

boomika_k@gwu.edu

Submitted to

Professor David Trott

Department of Computer Science

George Washington University

Submitted May 7th 2023

Abstract

The project aims to perform sentiment analysis on COVID-19 vaccine-related tweets using natural language processing techniques and supervised learning algorithms. The project classifies tweets into three categories: good, negative, and neutral using a sizable dataset of tweets gathered via the Twitter API. Three methods—TextBlob, VADER, and BERT models—were used to do the sentiment analysis. Data pre-processing, feature engineering, and data visualization were also used in this research to create word clouds and find the most prevalent phrases linked to certain emotions. The investigation also looked into the relationships between sentiment ratings and other characteristics including user location, tweet platform, word count, and average word length. The project's findings shed important light on the public's perception of the COVID-19 vaccine that can be used to shape public health policies and communication tactics.

TABLE OF CONTENTS

Introduction

The COVID-19 pandemic has brought significant disruption to the lives of millions of people worldwide, causing unprecedented changes in daily routines and social interactions. To curb the spread of the virus, governments and health organizations worldwide have launched vaccination campaigns. In this context, social media platforms, particularly Twitter, have emerged as crucial spaces for sharing opinions and discussing COVID-19 vaccines.

Analyzing the sentiment of tweets related to COVID-19 vaccines can offer valuable insights into public opinion, guide public health policies, and support effective communication strategies. Sentiment analysis on Twitter data has become an increasingly relevant research area, providing a comprehensive view of the public's perception of a particular topic. Sentiment analysis involves identifying the underlying sentiment or emotion in a piece of text.

In this project, we utilized a sentiment analysis model with transfer learning to classify tweets related to COVID-19 vaccines into positive, negative, and neutral categories. Our goal is to implement a machine learning model trained on a vast dataset of COVID-19 vaccine-related tweets collected via the Twitter API. We employed advanced natural language processing (NLP) techniques and supervised learning algorithms to train the model. Additionally, we performed data preprocessing, feature engineering, and data visualization techniques to generate word clouds and identify the most common phrases associated with each sentiment category. The results of this study provided valuable insights into the public's perception of COVID-19 vaccines, inform public health policies and communication strategies, and facilitate effective vaccination campaigns.

The results of this study provided valuable insights into the public's perception of COVID-19 vaccines, inform public health policies and communication strategies, and facilitate effective vaccination campaigns. By analyzing the sentiment of tweets related to COVID-19 vaccines, we can better understand public opinion and identify potential challenges in vaccine adoption. The outcomes of this project can help public health officials and policymakers create targeted strategies to promote vaccination and address vaccine hesitancy.

Literature Review

Several studies have been conducted on the sentiment analysis of COVID-19 vaccine-related discussions on social media, particularly Twitter. Three noteworthy studies are summarized below:

Qorib et al. (2023) conducted a study on COVID-19 vaccine hesitancy using sentiment analysis, text mining, and machine learning algorithms. The study analyzed 3 sentiment computation methods (Azure Machine Learning, VADER, and TextBlob) and 5 learning algorithms (Random Forest, Logistics Regression, Decision Tree, LinearSVC, and Naïve Bayes) with different combinations of three vectorization methods (Doc2Vec, CountVectorizer, and TF-IDF) to classify public sentiment into positive, neutral, or negative. The study concluded that combining Potter stemming and lemmatization increased model performances, and the TextBlob + TF-IDF + LinearSVC approach achieved the best performance in classifying sentiment with an accuracy, precision, recall, and F1-score of 0.96752, 0.96921, 0.92807, and 0.94702, respectively.

Liu and Liu (2021) analyzed English-language tweets related to COVID-19 vaccines between November 1, 2020, and January 31, 2021, to identify themes and temporal trends in the sentiment of COVID-19 vaccine-related tweets. The study applied the Valence Aware Dictionary and sEntiment Reasoner tool to calculate the compound score to determine whether the sentiment mentioned in each tweet was positive, neutral, or negative. The study concluded that the overall sentiment trend for COVID-19 vaccines on Twitter was positive, with spikes in negative sentiment appearing after the rollout of certain vaccines. The study also found that sentiment varied by country and region.

Shamrat et al. (2021) conducted sentiment analysis on Twitter tweets related to COVID-19 vaccines using NLP and supervised KNN classification algorithms. The study classified the sentiment of tweets related to Pfizer, Moderna, and AstraZeneca vaccines into positive, negative, and neutral categories. The analysis revealed that Pfizer showed 47.29% positive, 37.5% negative, and 15.21% neutral sentiment, Moderna showed 46.16% positive, 40.71% negative, and 13.13% neutral sentiment, and AstraZeneca showed 40.08% positive, 40.06% negative, and 13.86% neutral sentiment.

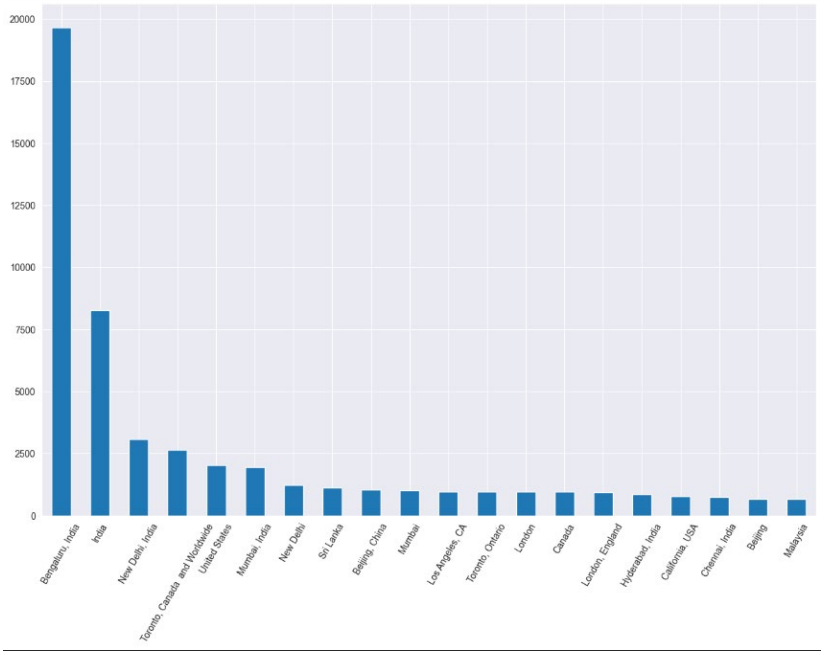
These studies demonstrate the potential of sentiment analysis to provide valuable insights into public opinion regarding COVID-19 vaccines, informing public health policies and communication strategies, and facilitating effective vaccination campaigns.

Methodology

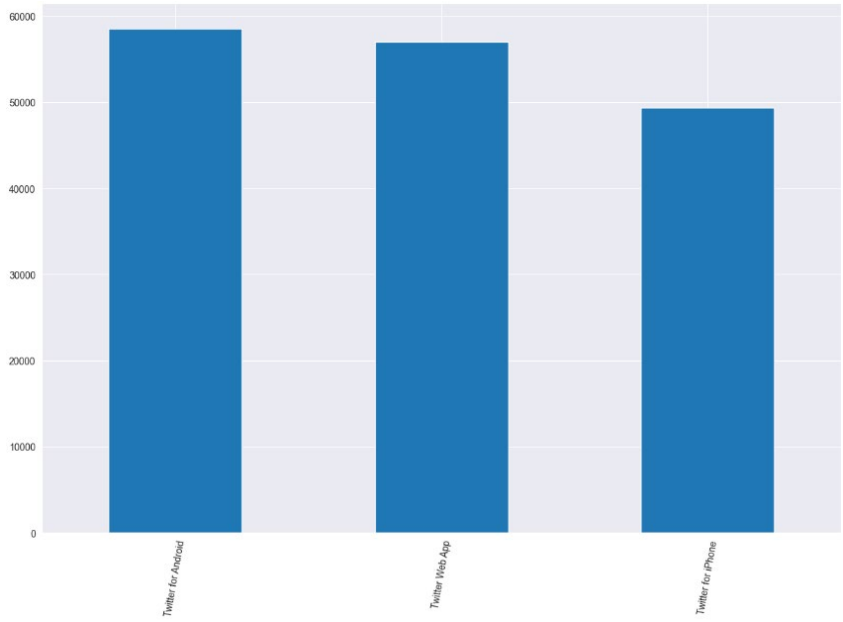
This project involves conducting sentiment analysis on tweets related to COVID-19 vaccines using Python code. The approach includes importing necessary libraries and datasets, data preprocessing, and two approaches to sentiment analysis: TextBlob and VADER.

The first step is importing the required modules, including os, numpy, pandas, re, matplotlib.pyplot, plotly, string, seaborn, nbformat, nltk, wordcloud, collections, and neattext. The script downloads and imports the 'vaccination_all_tweets.csv' and 'vaccination_tweets_trained.csv' datasets.

The data preprocessing step involves examining the dataset by displaying the head, tail, shape, and info of the 'vaccination_all_tweets.csv' dataset. Figure 1 displays a geographical and platform-wise distribution of tweets, providing a visual representation of the dataset for analysis. The 'text' column is converted to lowercase to standardize the data, and the neattext library is used to clean the data. The script removes hashtags, user handles, URLs, emojis, and special characters using the remove_hashtags(), remove_urls(), remove_userhandles(), remove_special_characters(), and remove_emojis() functions, respectively. Punctuation marks are removed using the remove_puncts() function. Stop words are also removed from the 'clean_data' column using the remove_stopwords() function. The script generates a word count using the Counter() function to get the most common words in the dataset.



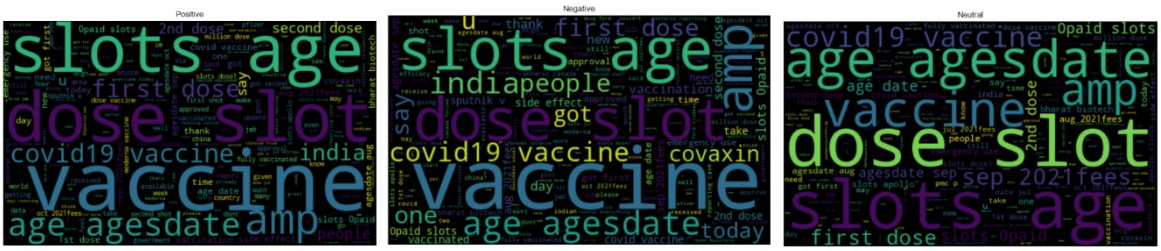
(a) Distribution of tweets by location



(b) Distribution of tweets by platform

Figure 1. Tweet distribution visualization

The same data preprocessing steps are applied to the 'vaccination_tweets_trained.csv' dataset.



(a) Positive *(b) Negative* *(c) Neutral*

Figure 2. Word cloud obtained from classification by TextBlob

The sentiment analysis step involves two approaches: TextBlob and VADER.

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. The reason we picked TextBlob for our analysis is because the library helps us train an already trained model for sentiment analysis on our own dataset and apply to a same context data. This is a core philosophy behind transfer learning.

VADER (Valence Aware Dictionary for Sentiment Reasoning), on the other hand, is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. It is available in the NLTK package and can be applied directly to unlabeled text data. It is a lexicon and rule-based feeling analysis instrument that is explicitly sensitive to suppositions communicated in web-based media. Because it is a rule based model, it is not possible to train in on a pre labeled dataset however, generating sentiments from it can help compare the results with TextBlob an provide valuable insights into the text.

Our TextBlob approach involves training the TextBlob model on the pre-labeled dataset from 'vaccination_tweets_trained.csv' using the NaiveBayesClassifier() function. The trained model is then applied to the 'clean_data' column in the 'vaccination_all_tweets.csv' dataset using the blob_fun() function to assign polarity and subjectivity to each tweet. The results are stored in a new 'results' column, and the tweets are separated into three categories based on sentiment: positive, negative, and neutral. Word clouds and bar graphs are generated for each category using the cloud_of_Words() and token_df_vis() functions. Figure 2 displays the word cloud generated based on the TextBlob sentiment analysis classification.

The VADER approach involves evaluating the accuracy of the VADER model on the pre-labeled dataset from 'vaccination_tweets_trained.csv'. The model is then applied to the un-labeled dataset from 'vaccination_tweets_unlabeled.csv' to predict the sentiment of each tweet. The results are stored in a new 'VADER_sentiment' column in the 'vaccination_tweets_unlabeled.csv' file.

Overall, this methodology involves a structured approach to sentiment analysis on COVID-19 vaccine-related tweets, including data preprocessing, two approaches to sentiment analysis, and data visualization to gain insights into public opinion and inform public health policies and communication strategies.

Results

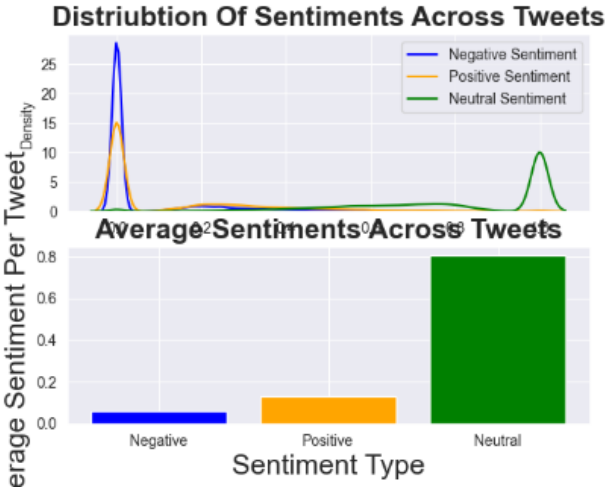


Figure 3. Distribution and average of sentiments across tweets

The project was executed on data obtained by Kaggle, which collected tweets from Twitter API. After data pre-processing and applying different supervised models, the resulting data were plotted to show the distribution of the sentiments across tweets.

On the pre-labeled dataset, the TextBlob model provides an accuracy of 72.91% while the VADER model has an accuracy of 47.61%. Due to this we consider TextBlob while performing the basic sentiment analysis. It returns us polarity, subjectivity and sentiment for the tweets.

Figure 3 shows the resulting distribution and average of sentiments across all COVID-19 vaccine-related tweets.

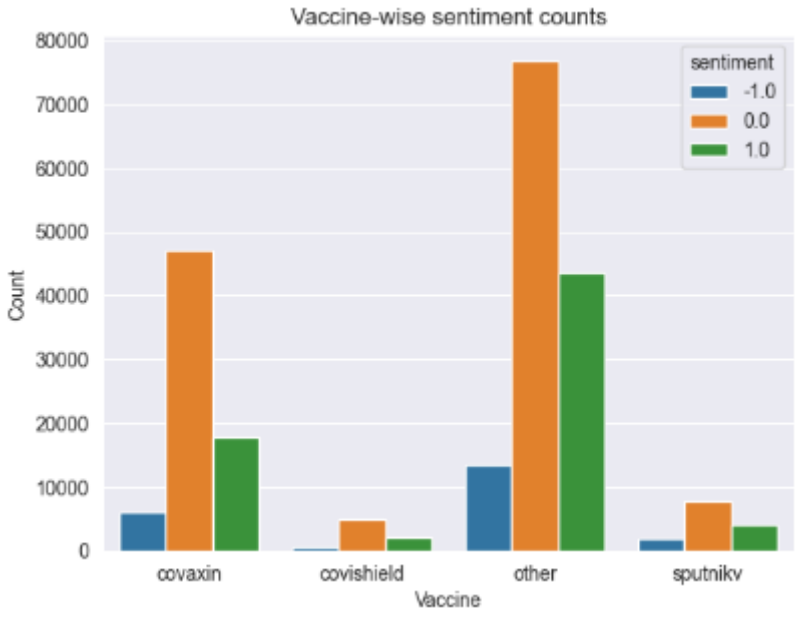
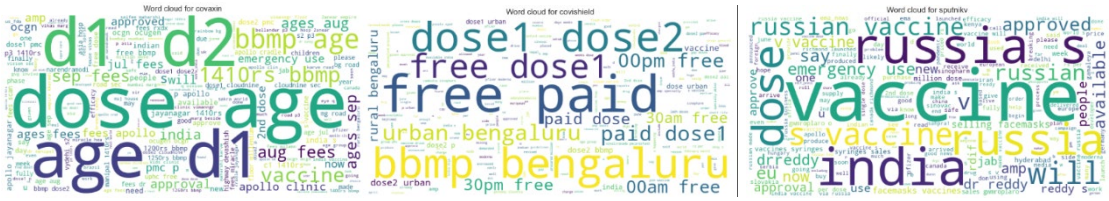


Figure 4. Vaccine wise sentiment count

Analysis was then conducted involving extracting data from vaccine-related tweets and analyzing the sentiment associated with each vaccine. The sentiment analysis involved grouping tweets by vaccine and assigning numeric values to sentiment values before generating

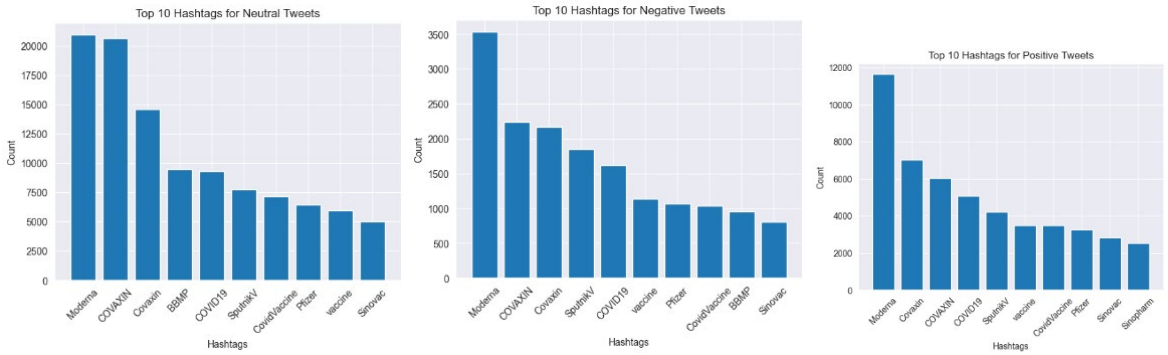
a bar plot for the sentiment counts of each vaccine. Figure 4 provides a visual representation of the sentiment counts for each vaccine.



(a) Covaxin *(b) Covishield* *(c) Sputnik*

Figure 5. Word cloud for tweets of popular vaccine brand

Additionally, word clouds were generated for each vaccine to visualize the most frequently occurring words. Figure 5 provides insights into the most frequent words used in tweets related to the most popular vaccines.



(a)Neutral *(b)Negative* *(c)Positive*

Figure 6. Hashtag analysis of tweets based sentiments

Finally, hashtag analysis was performed based on the sentiments. The resulting plot found that Moderna was the hashtag used most for all sentiments, followed by COVAXIN. Sinovac and Sinopham relatively have low hashtag frequency.

Overall, the analysis provides insights into the public sentiment toward COVID-19 vaccines. The analysis found that the sentiment towards the top vaccine for COVID-19 was mostly neutral, with the remaining majority of tweets being positive.

Conclusion

Our project's sentiment analysis code for Twitter data provides a powerful tool to monitor public opinion on any topic. The analysis indicates that most tweets related to COVID-19 vaccines are neutral, with a majority being positive and a smaller percentage being negative. The code is customizable for specific needs, such as analyzing sentiment on a particular brand or product. The generated visualizations and word clouds offer valuable insights into the most commonly associated sentiments and words with the topic of interest. This code is a valuable resource for businesses, policymakers, and researchers looking to track public opinion on social media.

References

Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Science Direct*, *212*(118715). doi:https://doi.org/10.1016/j.eswa.2022.118715.

Liu, S., & Liu, J. (2021). Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis. *Science Direct*, *39*(39), 5499–5505. doi:https://doi.org/10.1016/j.vaccine.2021.08.058.

Shamrat, F. M. J. M., Chakraborty, S., Imran, M., Muna4, J. N., Billah, M. M., Das, P., & Rahman, M. O. (2021). Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. *Indonesian Journal of Electrical Engineering and Computer Science*, *23*(1), 463–470. doi:http://doi.org/10.11591/ijeecs.v23.i1.pp463-470.

Gabriel Preda. (2021). COVID-19 All Vaccines Tweets [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2845240

Hassanhshah. (2021, October 17). *COVID vaccine sentiment and time series analysis*. Kaggle. Retrieved April 7, 2023, from https://www.kaggle.com/code/hassanhshah/covid-vaccine-sentiment-and-time-series-analysis

Amartyanambiar. (2021, September 11). *Covid-19 vaccine sentitmental analysis*. Kaggle. Retrieved May 7, 2023, from https://www.kaggle.com/code/amartyanambiar/covid-19-vaccine-sentitmental-analysis/notebook

Gpreda. (2021, March 20). *Explore vaccines tweets*. Kaggle. Retrieved May 7, 2023, from https://www.kaggle.com/code/gpreda/explore-vaccines-tweets