UNIVERSITÀ DEGLI STUDI
DI MILANO

Natural Language Processing and Sentiment Analysis: Tweets about covid-19 vaccination

Farnaz Sharbafi

MSc student in Data Science and Economics

Università degli Studi di Milano

Farnaz.sharbafi@studenti.unimi.it

Abstract

This project focuses on analyzing a dataset containing Twitter user information and tweet details to understand sentiment trends related to Covid-19 vaccines. Key columns considered in this analysis are the date of the tweet, tweet text, and the date the user created their account. The data underwent extensive cleaning, including the removal of mentions, hashtags, stop words, additional spaces, punctuation, and emojis. Polarity detection and sentiment analysis were conducted using two approaches: SentimentIntensityAnalyzer (VADER) and TextBlob. While SentimentIntensityAnalyzer, a rule-based approach, excels in handling informal language and social media text, TextBlob, a machine learning model, may perform better with formal language but struggles with emojis, slang, and emoticons. Results indicate that VADER identifies more negative tweets compared to TextBlob. Analysis also reveals trends in sentiment over time, the impact of user account age on tweet sentiment, and frequent words used in positive, negative, and neutral tweets. Additionally, sentiment analysis was conducted using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization combined with a Linear Support Vector Classifier (LinearSVC). This method demonstrated high effectiveness, achieving an accuracy of 86.43%. The findings suggest that tweets from older accounts are more prevalent, and that sentiment varies significantly based on the analysis approach used.

Keywords: Sentiment Analysis, Tweeter data, VADER, TF-IDF

Introduction

Dataset contains id, user_name, user_location, user_description, user_created, user_followers, user_friends, user_favourites, user_verified, date, text, hashtags, source, retweets, favorites, is_retweet columns. In this analysis date (date of tweet), text (tweet text) and user_created (user create date) columns are considered.

Data cleaning process is implemented. Words about mentioning another user on Twitter and hashtag sign are removed. Words are tokenized, stop words, additional spaces, punctuations and emojis are removed.

This project addresses the following key questions:

How do VADER and TextBlob compare in detecting sentiment from Covid-19 vaccine-related tweets, especially in handling informal versus formal language?
How does sentiment polarity vary over time as analyzed by VADER and TextBlob?
How does sentiment expression vary among users based on the year their accounts were created?
What are the dominant themes and sentiments expressed in positive, negative, and neutral tweets about Covid-19 vaccines?
How effective is TF-IDF vectorization combined with LinearSVC for sentiment classification of tweets?

Polarity Detection and sentiment analysis

Polarity detection was conducted using two sentiment analysis approaches: SentimentIntensityAnalyzer (VADER) from NLTK and TextBlob. SentimentIntensityAnalyzer relies on a pre-trained lexicon and a rule-based approach, while TextBlob uses a machine learning model for sentiment analysis. According to Govindappa and Channegowda (2022), both TextBlob and VADER offer significant features, each with unique advantages and limitations. TextBlob efficiently handles large volumes of text data with minimal computational overhead and is best suited for formal text. In contrast, VADER excels with text that includes slang and emojis due to its extensive sentiment lexicons, though it imposes a slightly higher computational overhead. For preprocessed datasets without informal language, TextBlob is generally more effective. Overall, SentimentIntensityAnalyzer handles informal language and social media text better due to its rule-based approach, while TextBlob performs better with formal language and nuanced contexts. Polarity scores greater than zero indicate positive tweets, scores equal to zero indicate neutral sentiment, and scores less than zero indicate negative tweets. Figure 1 demonstrates that VADER identifies more negative texts than TextBlob.

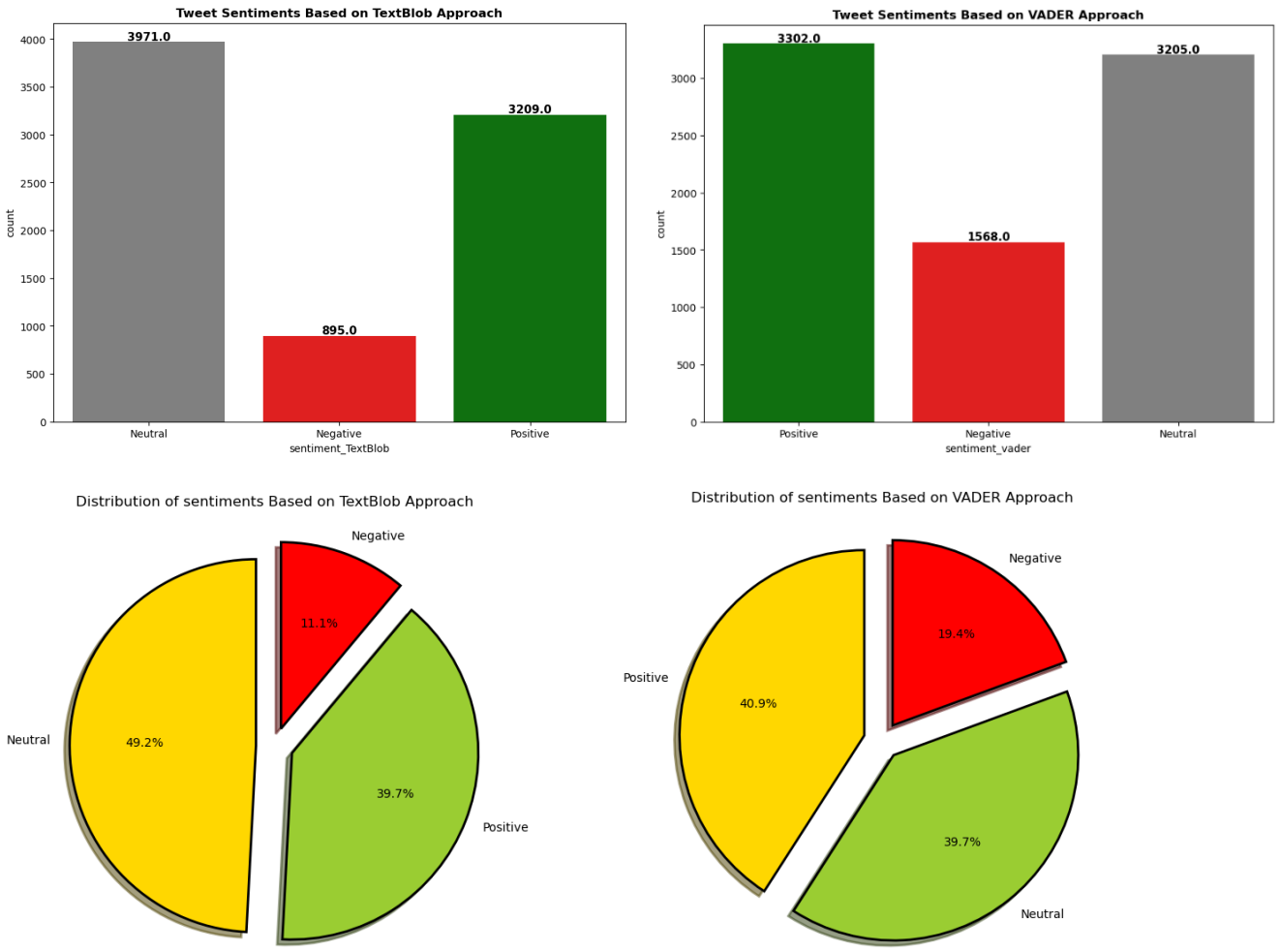


Figure 1. Number of tweets in each sentiment and distribution of sentiments based on VADER approaches

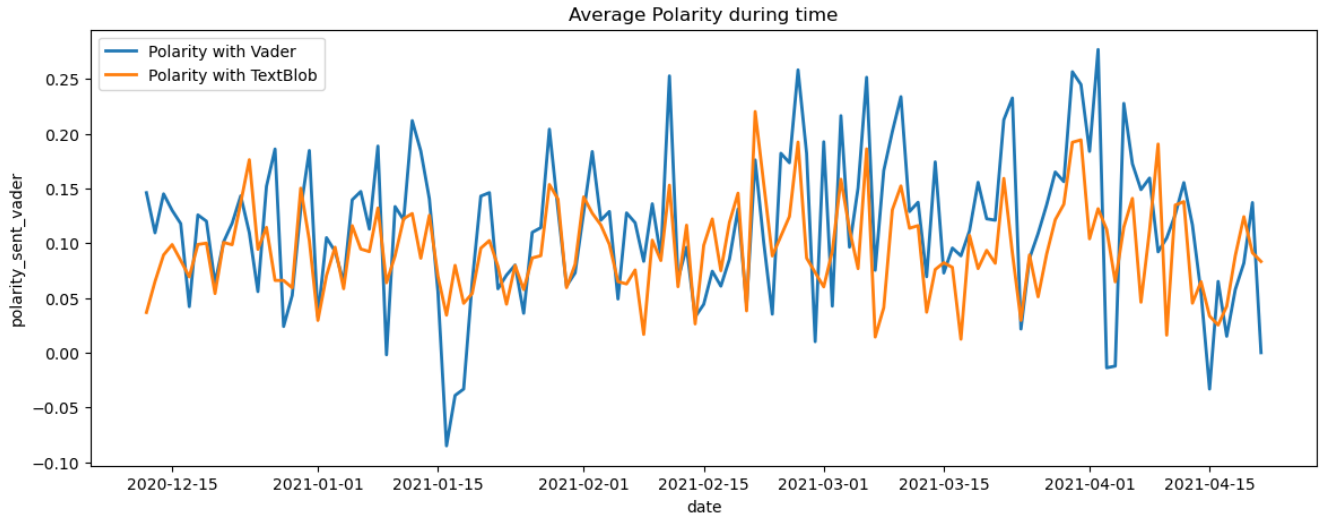


Figure 2. daily average of polarity with Vader and TextBlob approaches.

Polarity detection was conducted using two sentiment analysis approaches: VADER and TextBlob. As shown in Figure 2, there are notable differences in polarity between the two approaches during certain

periods. Since this project focuses on analyzing a series of non-formal tweets, the VADER approach is preferred for its better handling of informal language and social media text.

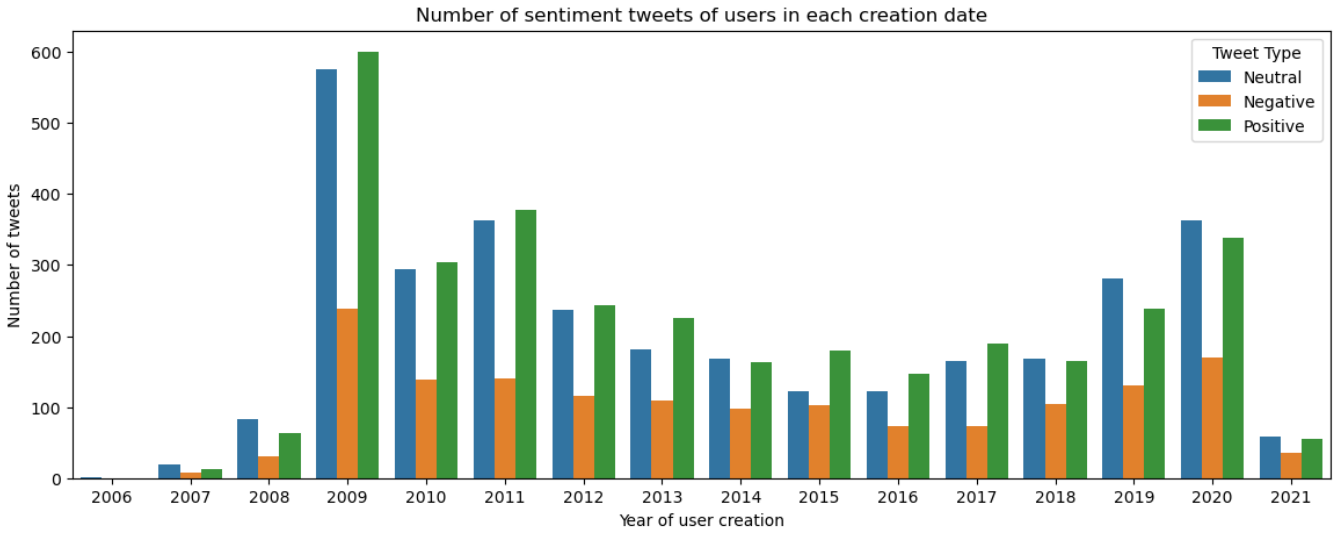
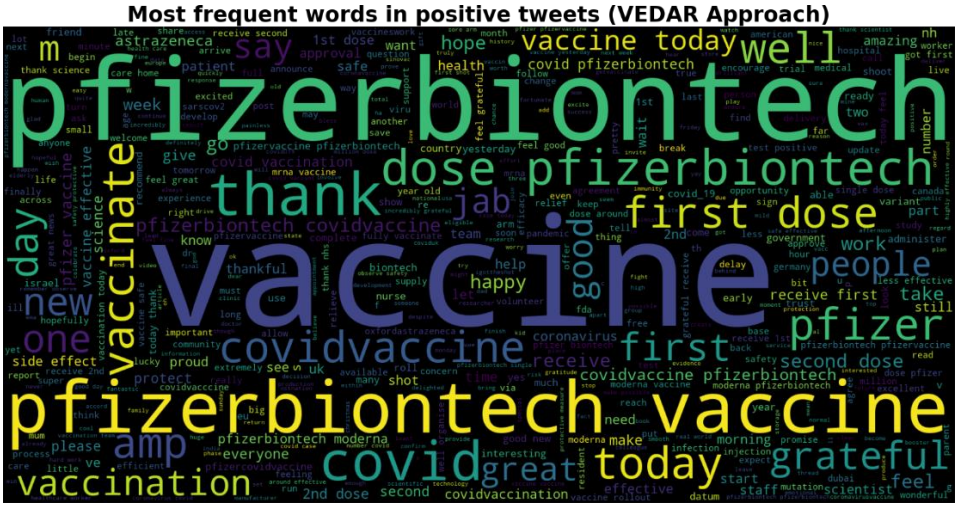


Figure 3. sum of sentiment tweets of users in each creation year.

Figure 3 reveals that most tweets are authored by users who created their accounts between 2009 and 2011. Additionally, it shows that the disparity between positive and negative tweets is more pronounced among older users, with older users posting a significantly higher number of positive tweets.

Frequent words in positive, negative, and neutral tweets

In this section, the frequency of words and adjectives is analyzed. Figure 4 presents a word cloud of frequent words, the top twenty frequent words (in a bar chart), the top twenty trigram words, and the top twenty frequent adjectives used in positive tweets. The analysis reveals that most of these words and adjectives pertain to explaining the effectiveness of the vaccine, providing instructions about safety measures, and announcing the receipt of the vaccine along with positive sentiments about it.



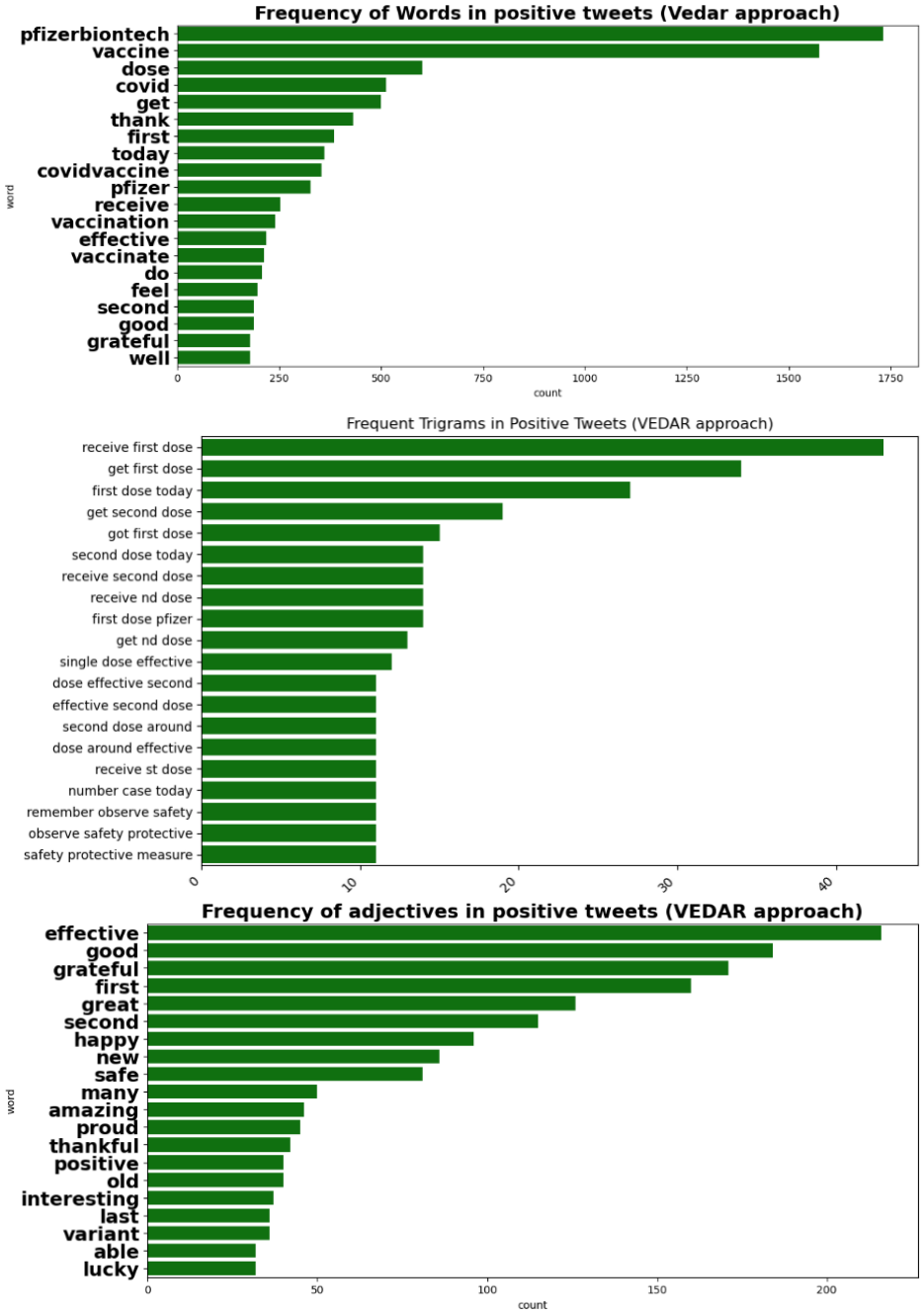
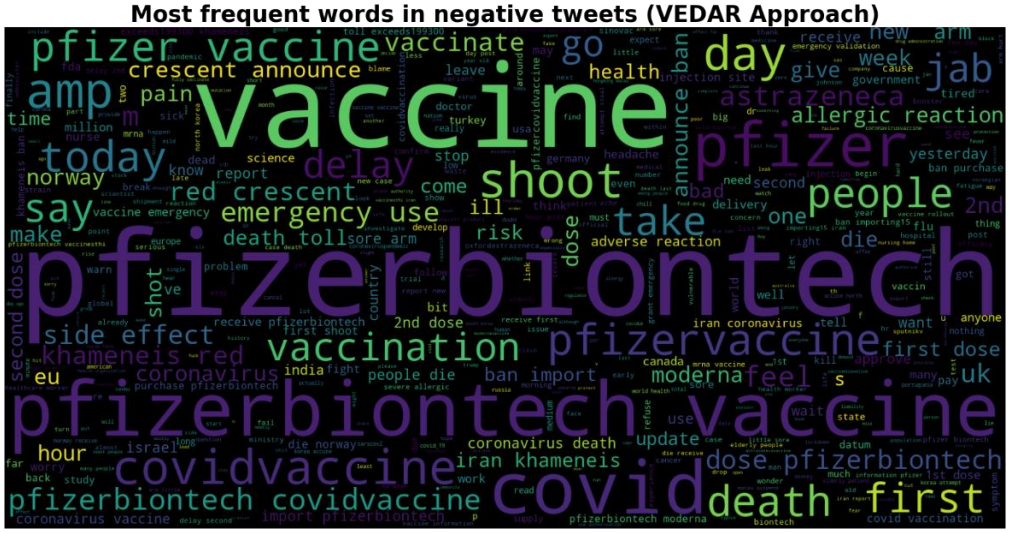


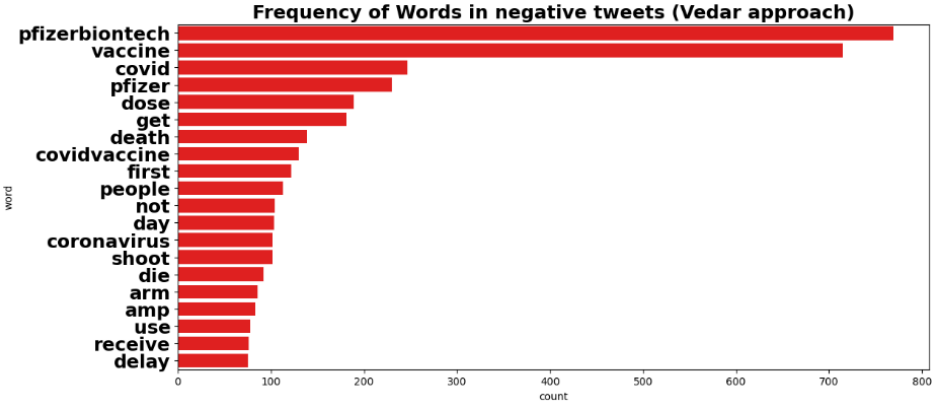
Figure 4. Frequent words and adjectives in positive tweets.

Figure 5 shows a cloud word of frequent words, top twenty frequent words (in bar chart), the top twenty trigram words, and also top twenty frequent adjectives used in negative tweets. We can understand that most of them are about explaining side effect of vaccine, vaccine import ban in Iran, and death toll information.

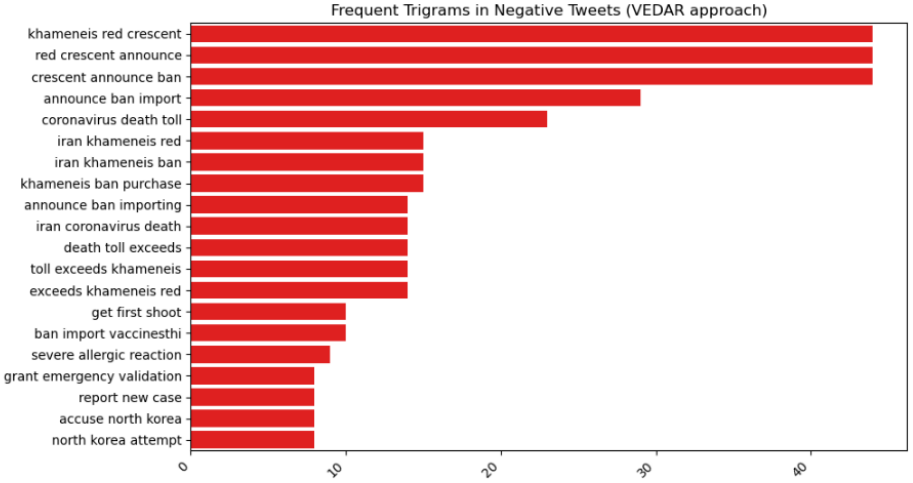
Most frequent words in negative tweets (VEDAR Approach)



Frequency of Words in negative tweets (Vedar approach)



Frequent Trigrams in Negative Tweets (VEDAR approach)



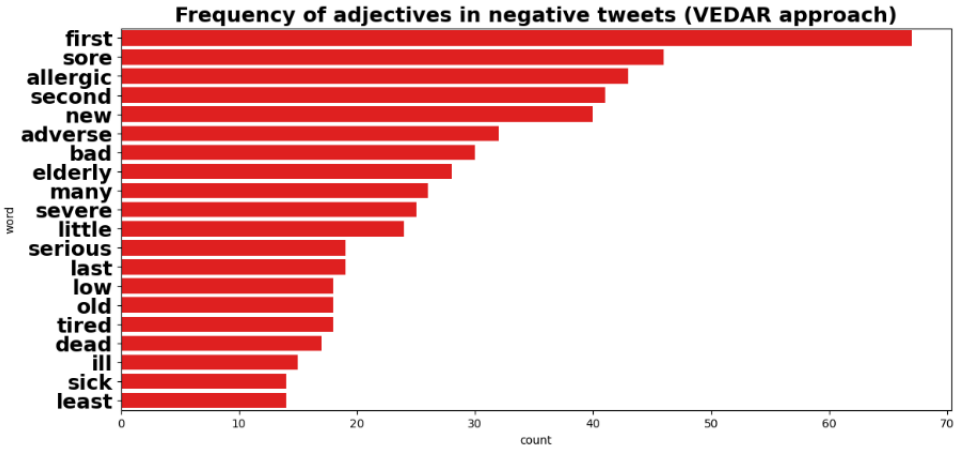
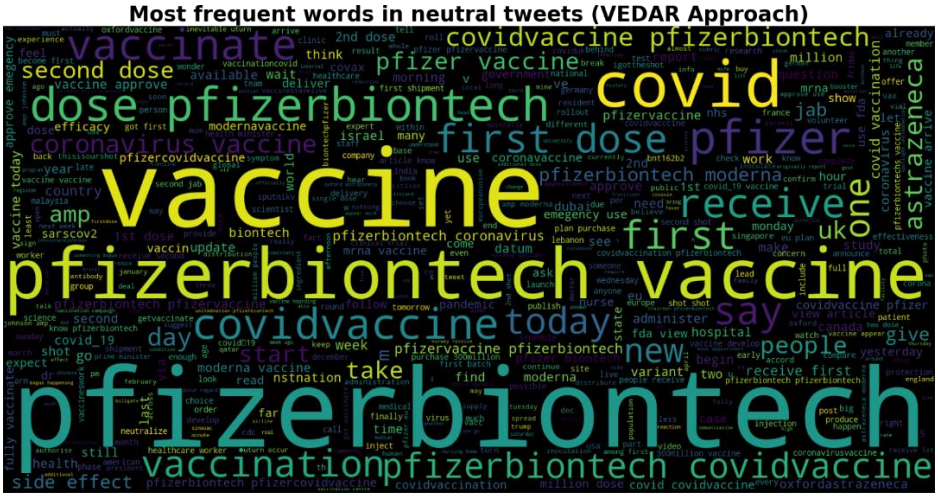
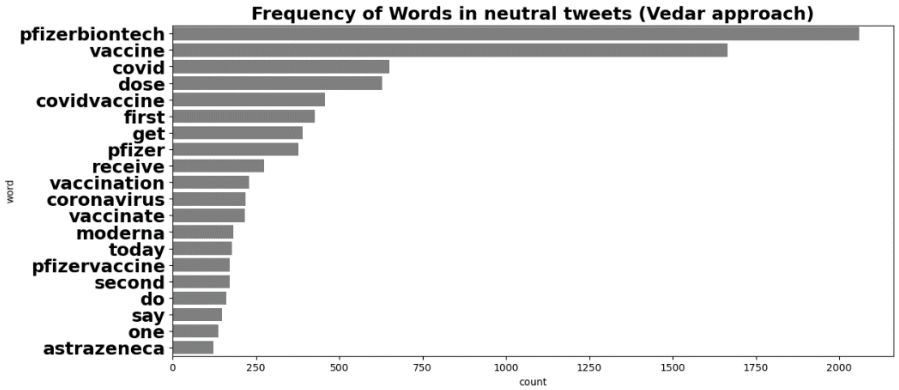


Figure 5. Frequent words and adjectives in negative tweets.

Figure 6 shows a cloud word of frequent words, top twenty frequent words (in bar chart), the top twenty trigram words, and also top twenty frequent adjectives used in neutral tweets. We can understand that most of them are about announcing news on Covid19 and vaccine.





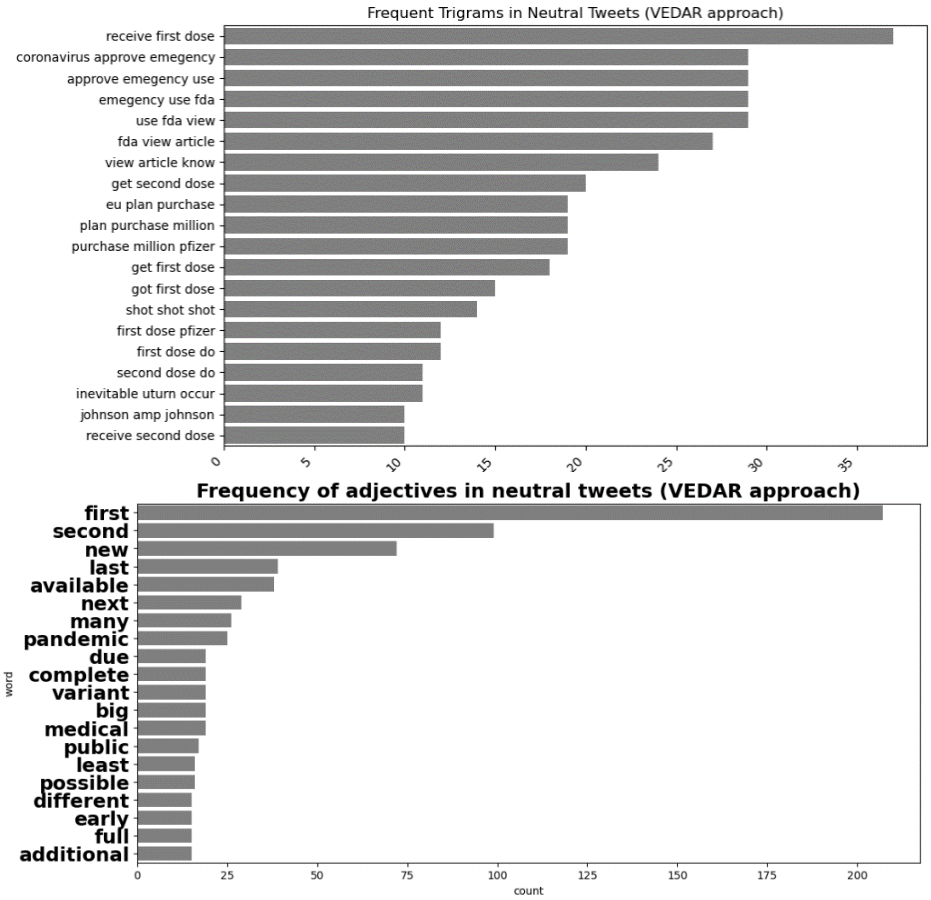


Figure 6. Frequent words and adjectives in neutral tweets.

Sentiment Analysis and Classification Using TF-IDF and LinearSVC

In this project, we conducted sentiment analysis on tweets by transforming the text data into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. The TfidfVectorizer from sklearn.feature_extraction.text was employed to convert the text data into a matrix of TF-IDF features. Subsequently, the dataset was split into training and testing sets using the train_test_split function from sklearn.model_selection, with 25% of the data allocated for testing and a random state of 42 to ensure reproducibility.

A Linear Support Vector Classifier (LinearSVC) was then implemented on the training data and used to make predictions on the test data. The analysis demonstrated the effectiveness of combining TF-IDF vectorization with a LinearSVC model for sentiment classification of tweets, achieving an accuracy of 86.43%, the approach effectively captured nuances in sentiment across Negative (79% precision, 69% recall), Neutral (84% precision, 95% recall), and Positive (92% precision, 85% recall) sentiment categories.

The detailed evaluation metrics, including precision, recall, and f1-score, indicated good performance across different sentiment classes that are shown in figure 7. The confusion matrix and classification report provided further insights into the model's performance. For example, the confusion matrix showed how well the model differentiated between negative, neutral, and positive tweets. The

precision, recall, and f1-score for each sentiment class highlighted the model's ability to correctly identify each class, with higher values indicating better performance.

This analysis confirms the suitability of using TF-IDF vectorization and LinearSVC for sentiment analysis of tweets, effectively capturing the nuances in the sentiment expressed in social media text.

	precision	recall	f1-score	support
Negative	0.79	0.69	0.74	367
Neutral	0.84	0.95	0.90	833
Positive	0.92	0.85	0.88	819
accuracy			0.86	2019
macro avg	0.85	0.83	0.84	2019
weighted avg	0.87	0.86	0.86	2019

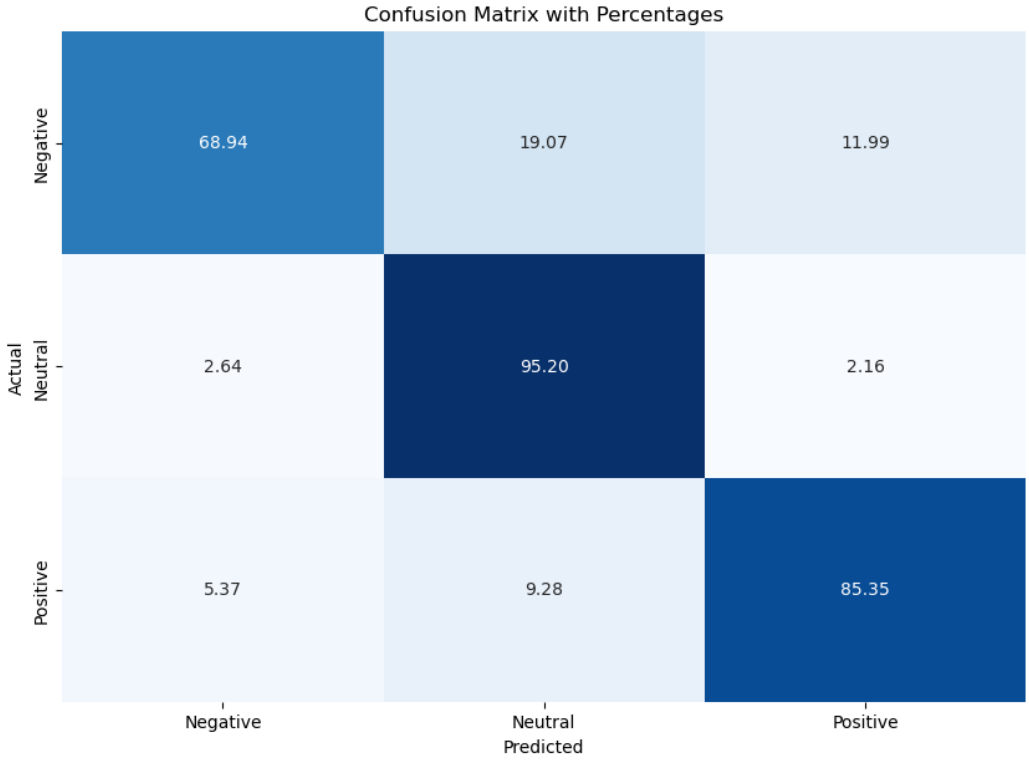


Figure 7. Classification result using TF-IDF and LinearSVC

Conclusion

The sentiment analysis of Covid-19 vaccine-related tweets using both VADER and TextBlob provides insightful distinctions between the two approaches. VADER's ability to handle informal language and social media-specific elements makes it a more suitable choice for analyzing non-formal tweets, as demonstrated by its higher detection of negative sentiment. The analysis highlights that user accounts created between 2009 and 2011 are the most active, and these older accounts show a greater gap

between positive and negative sentiments. The frequent word analysis in tweets shows that positive tweets are generally about the effectiveness of vaccines and safety measures, while negative tweets focus on side effects and vaccine-related issues. Neutral tweets are primarily used for news announcements.

Additionally, employing TF-IDF vectorization with a Linear Support Vector Classifier (LinearSVC) demonstrated robust performance in sentiment classification. Achieving an accuracy of 86.43%, the approach effectively captured nuances in sentiment across Negative (79% precision, 69% recall), Neutral (84% precision, 95% recall), and Positive (92% precision, 85% recall) sentiment categories.

These insights can help public health officials and policymakers understand public sentiment and address concerns related to Covid-19 vaccines more effectively.

References

Govindappa, Jalaja, and Kavitha Channegowda. "Analyzing sentiment dynamics from sparse text coronavirus disease-19 vaccination using natural language processing model." *International Journal of Electrical and Computer Engineering (IJECE)* 12.4 (2022): 4054.