Forecasting Twitter Engagement: Insights and Predictions

Introduction

This project focuses on analyzing Twitter data to understand public sentiment related to COVID-19, particularly about the vaccines and different variants. Using PySpark, various data processing and machine learning techniques were applied to derive insights from tweet content, user engagement, and geographical trends, which can be valuable for public health communication in policy-making and understanding public opinion during the pandemic.

Dataset

This dataset holds twitter covid 19 tweets of 2023 and the discussions about vaccines. The dataset comprises ~60000 records, each representing a tweet from Twitter. Each record contains details such as the tweet's creation timestamp, unique identifier, full text, user information (including screen name and location), engagement metrics (retweet and favorite counts), and metadata (including language and source)

Sentiment Analysis of COVID-19 Discussions

We are analyzing the sentiments expressed within a collection of tweets. By employing a sentiment classification function, we aim to categorize each tweet as either Positive, Negative, or Neutral based on the polarity of its text. Through subsequent DataFrame processing and data aggregation, we seek to understand the distribution of sentiments among the tweets, providing insights into the overall sentiment trends within the dataset. Ultimately, the visualization of sentiment distribution via a bar plot (fig.1) allows for a clear and intuitive representation of the sentiment landscape within the tweets.

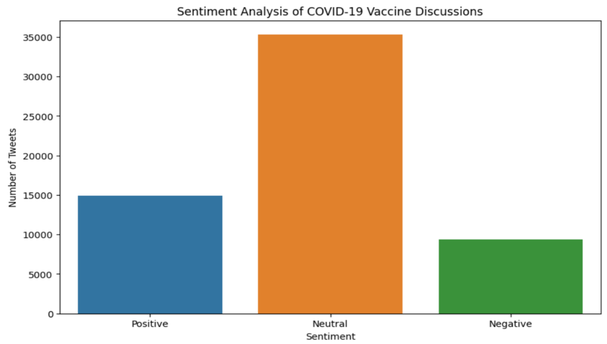

Figure 1: **Sentiment Analysis of COVID-19 Discussions**

Public Sentiment towards COVID-19 Variants

We are analyzing the sentiment expressed in tweets related to specific COVID-19 variants, such as Delta and Omicron. Through visualization(fig2), we seek to understand the distribution of sentiment towards each COVID-19 variant, highlighting any differences in public sentiment towards these variants. This analysis can provide insights into how the public perceives and responds to different COVID-19 variants.

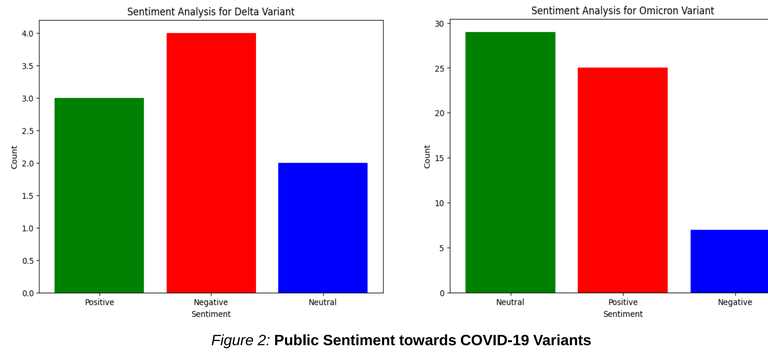

Figure 2: **Public Sentiment towards COVID-19 Variants**

Geographic Trends of COVID-19 Concerns

We are analyzing the geographic trends in COVID-19 concerns by extracting location data. The extracted locations are grouped to count the number of tweets per location. Dynamic thresholds for tweet counts are then calculated to categorize locations based on tweet volume, allowing us to identify areas of varying levels of COVID-19 discussion intensity. This analysis provides insights into the distribution of COVID-19-related discussions across different regions(fig3), helping to understand regional variations in public sentiment, awareness, and engagement with the pandemic.

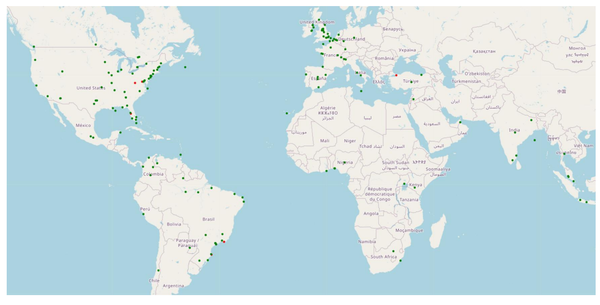

Figure 3: **Geographic Trends of COVID-19 Concerns**.

Twitter User Engagement Prediction

We are predicting user engagement, measured by retweet counts, based on features extracted from tweets. Textual and numerical features are extracted from the tweets using various techniques such as Tokenizer, StopWordsRemover, HashingTF, and IDF. A RF regressor is then trained using these features and engagement features (followers, friends counts). Ultimately, this analysis helps in understanding the factors influencing user engagement with tweets and in developing predictive models to forecast user interaction on social media platforms.

Below are the evaluation metrics:

R-squared (R2): 0.9649063044650737

Mean Absolute Error (MAE): 121.52988422575979

Finding the Big Talkers who are Verified Users

This task involves identifying and ranking verified users on social media platforms who are considered "Big Talkers" based on their influence in discussions related to COVID-19 (fig4). The data is filtered to include only verified users, and an influence score is calculated for each user, considering metrics such as followers and engagement. Subsequently, the data is aggregated by user, and influencers are ranked based on their total influence score. This analysis helps in understanding the key players shaping discussions and opinions on social media platforms during the pandemic.

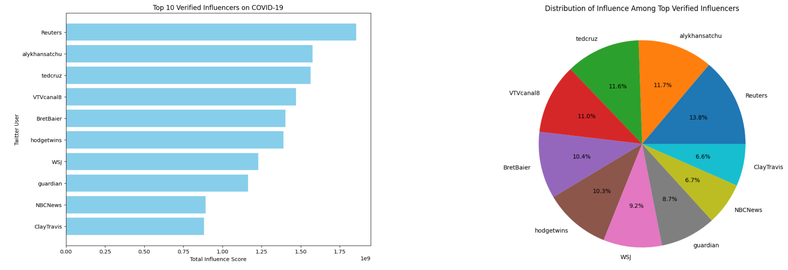

Figure 4: **Finding the Big Talkers who are Verified Users**

CONCLUSION

This analysis outlines the systematic approach used to extract, process, and analyze Twitter data for insights into public sentiment on COVID-19. The findings can aid in understanding public opinion and potentially guide public health strategies.