Sentiment Analysis of Pfizer Vaccine Discussions on Social Media: A Time Series Approach

By QUOC VIET LE
University of Kansas

I. Introduction. The COVID-19 pandemic has had a profound impact on global health and society, leading to the development and authorization of several vaccines for emergency use worldwide. However, achieving high vaccination rates has been hindered by vaccine hesitancy and skepticism, which can impede efforts to control the pandemic. Social media has become a significant platform for sharing information and opinions about vaccines, with public sentiment on social media having a considerable impact on vaccination rates. As one of the most widely administered vaccines globally, understanding public sentiment towards the Pfizer-BioNTech COVID-19 vaccine is crucial for achieving high vaccination rates and combating the pandemic. This project aims to investigate the evolution of public sentiment towards the vaccine on social media by analyzing a large dataset of tweets collected on Kaggle. VADER sentiment analysis will be used to obtain sentiment scores, and time series analysis will be employed to analyze how public sentiment has changed over time in response to various events and news related to the vaccine. The results of this study can provide valuable insights into the factors that influence public sentiment towards the vaccine and inform decision-making for healthcare professionals, policymakers, and pharmaceutical companies.

II. Methodology.

A. Data preprocessing and cleaning. The initial dataset contained many tweets and included various types of noise such as emojis, special characters, and URLs. To prepare the dataset for analysis, several preprocessing steps were performed, including removing duplicate entries, removing retweets, removing non-English tweets, and removing URLs and special characters. The remaining tweets were then analyzed using the VADER sentiment analysis tool to obtain sentiment scores, which were divided into three categories: positive sentiment, negative sentiment, and neutral sentiment. To achieve this, the text was tokenized and passed through the *polarity_scores()* function of the *SentimentIntensityAnalyzer* object from the Natural Language Toolkit (NLTK). The sentiment scores were then extracted and added to separate columns in the dataset. Additionally, to make the scores comparable across different tweets, a small value of 10^{-6} was added to each score. Finally, the 'sentiments' column was dropped from the dataset.

B. Exploratory data analysis. To gain insights into the sentiment trends of the dataset, we performed exploratory data analysis (EDA) using visualizations. First, we resampled the dataset to a frequency of 1 day. Then, we plotted the sentiment scores for each day using line charts. The sentiment scores were divided into three categories: positive sentiment, neutral sentiment, and negative sentiment.

Figure 1. Sentiment Scores toward Pfizer-BioNTech COVID-19 Vaccine from Dec 15, 2020 to Apr 15, 2021

The graph of sentiment scores of tweets collected over time reveals interesting trends (Figure 1).

Firstly, it is observed that both the positive and negative scores decrease towards the end of the data collection period. This suggests a decrease in the overall sentiment polarity of the tweets in the dataset towards the end of the period. However, it is also noted that the trend of the positive and negative scores over the

entire period is almost constant. This could mean that although the scores fluctuate, the overall sentiment polarity of the tweets remains relatively stable.

Secondly, the trend of the neutral sentiment scores over the entire period shows a slight upward trend, indicating that tweets with a neutral sentiment are becoming more prevalent. This trend, coupled with the relatively constant trend of the positive and negative scores, suggests that the overall sentiment polarity of the tweets in the dataset is becoming less extreme and more neutral.

Overall, the graph provided valuable insights into the patterns and trends of sentiment scores in the tweet dataset and helped to inform further analysis and interpretation of the data.

We conducted a further analysis of the positive sentiment score by removing the trend using polynomial regression. We fitted multiple polynomial regression models of increasing degrees (up to 10) and plotted the residual process against time.

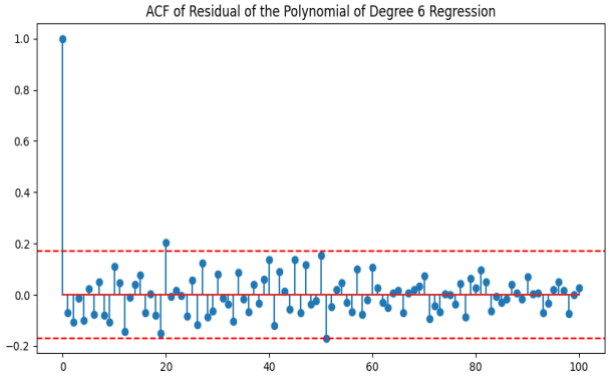

Figure 2. ACF plot of residual process of positive sentiment score

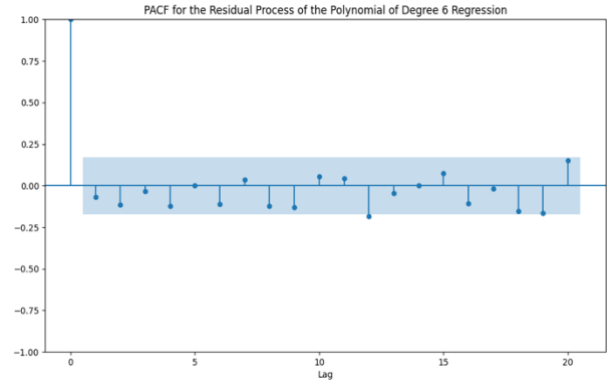

Figure 3. PACF of residual process of positive score

Our analysis involved examining the residual process's autocorrelation using the autocorrelation function (ACF) and partial autocorrelation function (PACF). We observed a significant cut-off at lag 20 in the ACF and lag 12 in the PACF in each polynomial regression. These findings suggest that an ARMA model with an autoregressive order of 12 and a moving average order of 20 may be appropriate for predicting the data trend. We used the Akaike information criterion (AIC) to select the best-fit model for predicting the data trend further. We fitted each residual process to a polynomial trend, which resulted in multiple potential models with varying degrees of complexity. We then calculated the AIC score for each model and selected the one with the lowest AIC score, considering both the goodness of fit and model complexity. We found that the polynomial of degree 6 regression was the best choice for ARMA(12,20). Furthermore, we examined the ACF and PACF plots to provide more insight into the residual process's correlation. The ACF plot (Figure 2) demonstrated significant correlations between residuals for most lags, with lag 20 falling outside of the confidence interval, indicating a significant dependence on the last error. Conversely, the PACF plot (Figure 3) displayed a significant cut-off at lag 12, highlighting the impact of the last 12 errors on the current value.

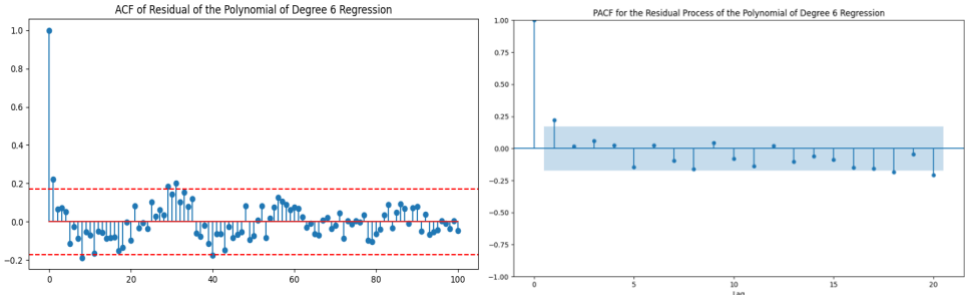

Figure 4. ACF and PACF of negative score of polynomial of degree 6 residual

After conducting EDA on the negative score data through visualization and ACF and PACF analyses, we observed that the ACF exhibited a tail-off pattern that fell within the confidence interval, and the first lag cutting off the confidence interval was at lag 8. Therefore, we determined that the MA(8) pattern was applicable. Additionally, the PACF also showed a tail-off pattern and nearly fell within the confidence interval. However, due to the high correlation of lag 1 to the last value, we still considered lag 1 in the PACF analysis. We evaluated the residual processes of polynomial fits for both ARMA(0,8) and ARMA(1,8) models and computed their respective AIC. Afterward, we determined that the ARMA(0,8) polynomial of degree 6 regression provided the lowest AIC score. We further examined the ACF and PACF plots to provide more insight into the residual process's correlation (Figure 4).

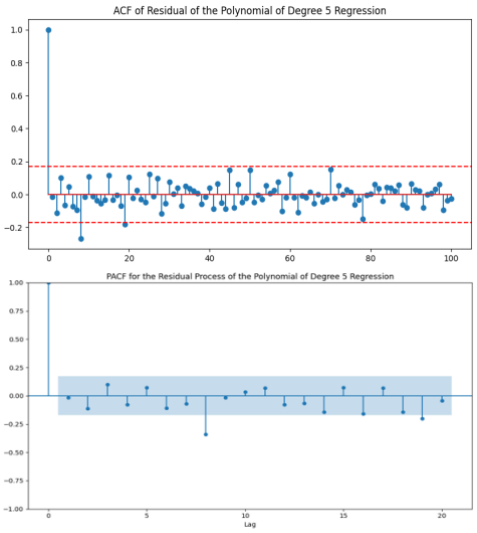

Figure 5. ACF and PACF of residual process of neutral score

Likewise, we discovered that the residual process of the neutral score could be fitted using a polynomial of degree 5 with ARMA(8,8) model, which also yielded the lowest AIC. We further analyzed the residual process by examining its ACF and PACF, providing additional insights into the neutral score data (Figure 5). With these models established, we can proceed with predicting future outcomes for the upcoming month (Figure 6).

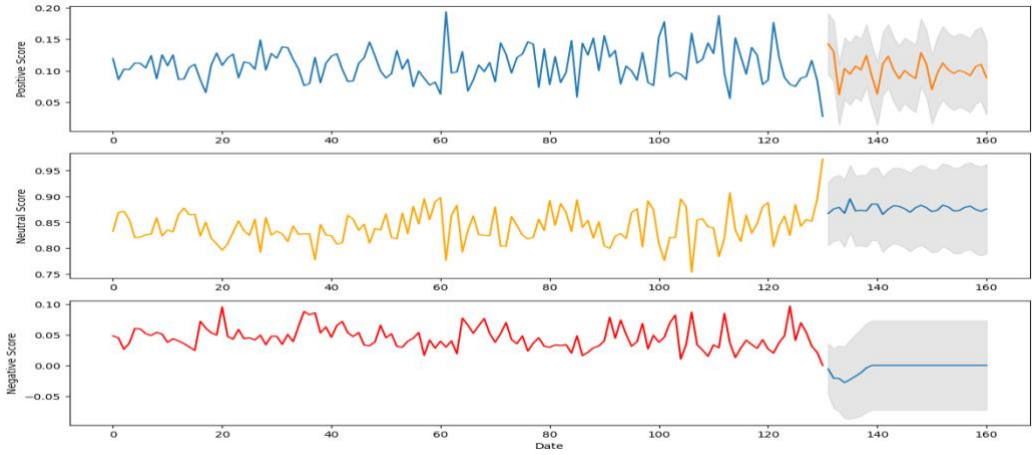

Figure 6. The plots of sentiment scores and their prediction within 95% confidence interval in May, 2021

III. Discussion. Upon analyzing the resampled dataset in dates rather than in minutes, we observed a clear discontinuity in scores occurring on April 16th, 2021 (at the point of 131) since we resampled the data to a frequency of 1 day and take the mean of sentiment scores within each interval. We utilized the ARMA models we established to predict future outcomes, with the different colored regions on the graphs representing these predictions and the gray regions indicating the 95% prediction intervals. Our findings indicate a positive trend in the acceptance of the Pfizer-BioNTech COVID-19 vaccine, with an increase in positive scores and a decrease in both negative and neutral scores over the next few days. Despite some fluctuations in positive scores, the overall trend remains promising. Furthermore, we observed a decline in neutral scores, and a significant decreasing trend in negative scores. These results provide evidence that people are gradually accepting the vaccine.

IV. Conclusion. Our analysis indicates that the Pfizer-BioNTech COVID-19 vaccine is gaining acceptance amongst the population. This study showcases the effectiveness of ARMA models in predicting trends and outcomes and provides valuable insights into the attitudes of the public towards the COVID-19 vaccine.

V. Reference.

[1] Brockwell, P. J., & Davis, R. A. (2016). Introduction to Time Series and Forecasting. Springer.

[2] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. Vader Sentiment Analysis. Retrieved from https://tjzhifei.github.io/resources/NLTK.pdf