Pfizer Tweets Sentiment Analysis

ZHANG HUIYAN #21436576
LIANG JINGBING # 21419728
YUAN YICHEN # 21421986

Abstract

As the number of deaths caused by the covid-19 pandemic continued to rise, Pfizer was developed to prevent infections. Hence, the goal of this data project is to figure out people's attitudes via sentiment analysis with 6 different modeling with the Pfizer twitter dataset from Kaggle.

1. Research Question

RQ1: What are people's attitudes about vaccinations?

RQ2: Which classification modeling can achieve the optimal results?

2. Methodology

2.1 Data Overview

The datasets contain 11,020 rows and 16 columns. However, some of the features in this dataset seem meaningless to our analysis. In this data project, we only picked up user locations, dates, and hashtags for exploratory analysis while the text for further text sentiment analysis.

2.2 Data Pre-processing

The first step in data preprocessing was to inspect the graph for missing values. The graph showed that the dataset contains no missing values. Second, we need to clean up the text to avoid noise. To remove punctuations, hashtags, URLs, digits, and lowered items, we used the basic gsub function. Before starting the modeling, we used the tidytext [1] packages to plot positive and negative word token distributions. The most positive words were "protect" (272) and "hope" (128), while "shoot" (536) and "grate" (188) were the most negative in Figure 1.

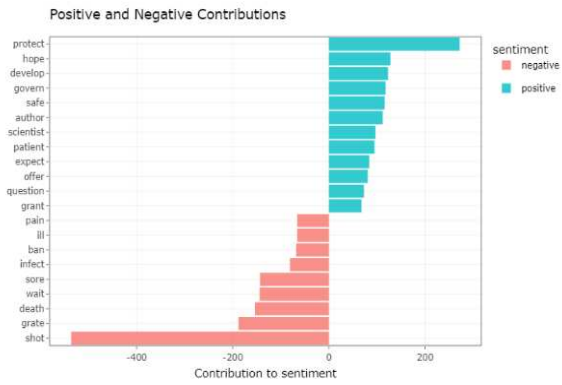

Figure 1 Top 15 positive and negative contributions

Before beginning the sentiment analysis, a new variable called mysentiment was created in sentiment analysis using nrc dictionary. In this case, we classified the term as positive if the positive score was greater than the negative score, and vice versa. Figure 2 showed that the number of positives (5,845) was greater than the number of negatives (5,175).

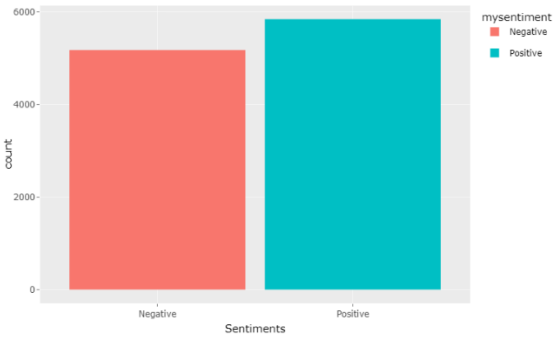

Figure 2 Positive and negative sentiments distributions

In addition to text preprocessing, we plotted a location graph of Twitter users. Figure 3 showed that Malaysia (168), London, England (142), and India(127) were in the top three. 4 showed that the number tweets continued to fall after the winter peak pandemic period though with a brief spike in the autumn of 2021. This trend closely corresponded to covid 19 virus active trend.We also explored the hashtags network of feature concurrences. Figure 5 showed that the majority of vaccines usually agreed with the covid and other vaccination names.

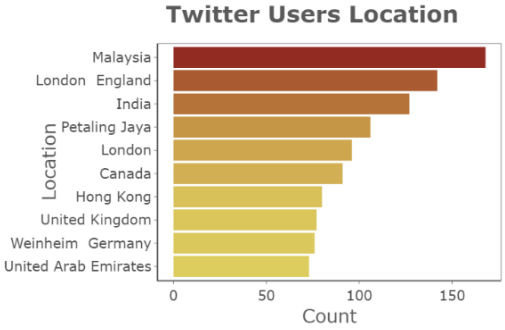

Figure 3 Top10 user locations

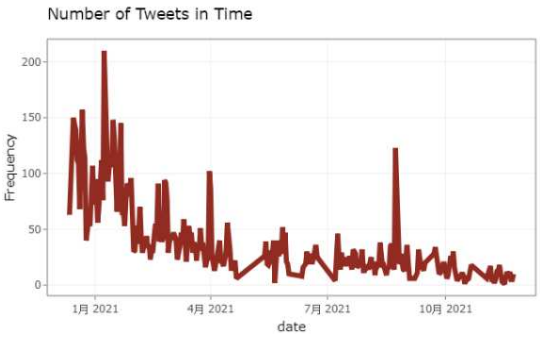

Figure 4 Number of Tweets in Time

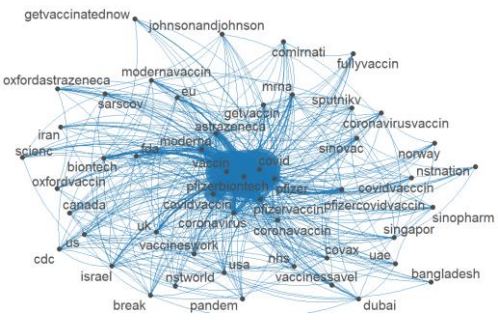

Figure 5 Hashtags concurrence networks

To have a brief glance at the positive and negative words word, we started by creating the corpus and then tokenizing it by removing unnecessary symbols and stop words. Then, we converted them to dfm after lemmatization and word stem. Figure 6 and Figure 7 showed that "covid" and "vaccine" were the most frequently used words in both positive and negative contexts. We also created a top 20 frequency graph as well as a cumulative frequency one. Figure 8 showed that the most frequently used words were vaccine and biotech. The cumulative frequency graph's increase rate slowed after the frequency order reached 750, as shown in Figure 9 .

Figure 6 Positive wordcloud

Figure 7 Negative wordcloud

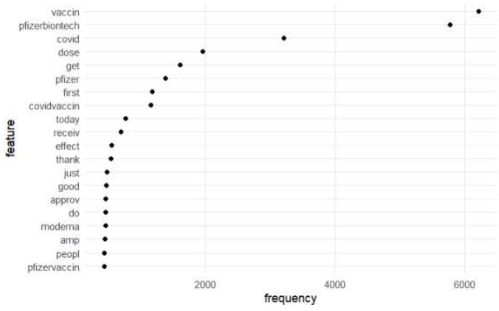

Figure 8 Top 20 frequent terms

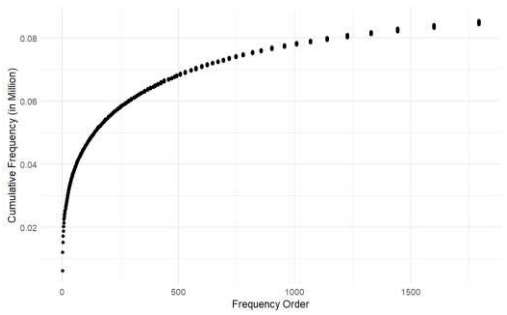

Figure 9 Cumulative sum frequency

3. Results

3.1 Train Test Split

Prior to modeling construction, the corpus was divided into 70% for the train set and 30% for the test set. Six different modelings were used because the target variable mysentiment was categorical.

3.2 Naïve Bayes

Here, we chose Naïve Bayes as our baseline model since it was easy to explain. The top three positive words "vaccine", "pfizerbiontech", and "covid" did not appear to convey the correct positive sentiment, whereas the top three negative words "toll", "avoid", and "boycott" did.

3.3 Ridge Regression

Then, ridge regression was selected because it could reduce model complexity by applying an L2 penalty. The top three positive words "vaccin", "grate", and "good" shown in Figure 10 appeared to make sense, while the top three negative words "shoot", "pandem", and "ill" made sense too.

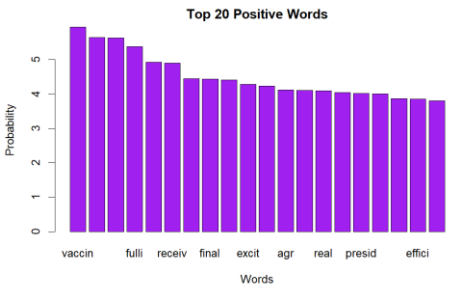

Figure 10 Top 20 Positive words in Ridge Regression

3.4 Support Vector Machine

The Support Vector Machine was chosen because it can easily handle large amounts of high-dimensional data. The model we chose has an accuracy rate of 0.79, which was slightly higher than Ridge Regression.

3.5 Random Forest

Random Forest was chosen because it could reduce overfitting and variance. Unlike the traditional method, we chose h2o packages to save time. The ntrees parameter was set to 100, while the nfolds parameter was set to 5. The variable importance graph, however, could only tell you the overall variable importance, not the top positive and negative. Figure 11 showed that the most important words were "vaccin", "receiv", "shoot," "grate," "fulli", and "covid".

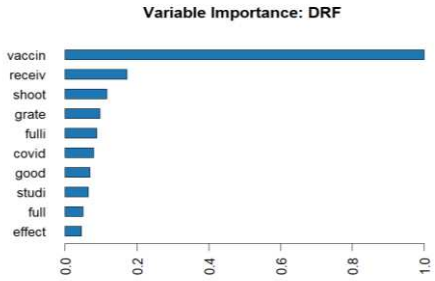

Figure 11 Random Forest Variable Importance

3.6 Gradient Boosting

Gradient Boosting was chosen because it was easy to reduce bias and was less prone to outliers. The variable importance graph results were nearly identical to the Random Forest results.

3.7 Deep Learning

To achieve the best classification results, Deep Learning with RectifierWithDropout pattern was used.

Figure 12 shows that the most important words were "vaccin", "bp", "plea", "malaysian", "borisjohnson", and "chairman".

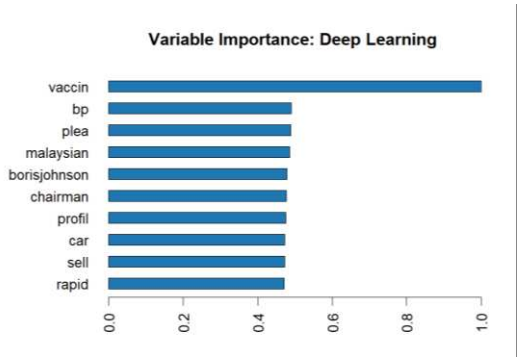

Figure 12 Deep Learning Variable Importance

3.8 Evaluation Metrics

3.8.1 ROC Curve

Figure 13 showed that ridge regression had the highest AUC of 0.866, followed by Support Vector Machine at 0.86. Deep Learning had the lowest AUC value.

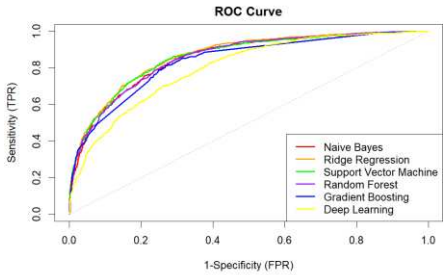

Figure 13 ROC Curve

3.8.2 F1 Score

Figure 14 showed that Gradient Boosting had the highest F1 score of 0.806, followed by Random Forest (0.798) and Deep Learning (0.796). In terms of F1 score, naive Bayes performed the worst.

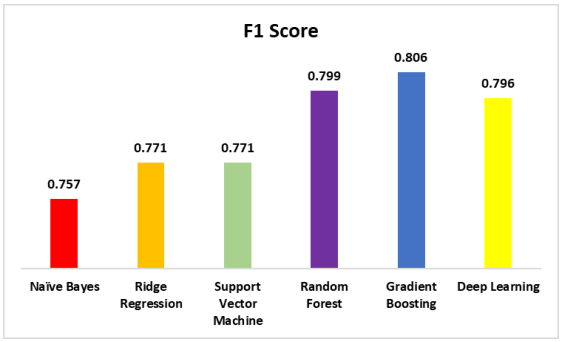

Figure 14 F1 Score

4. Conclusions

Most people had positive attitudes toward Pfizer vaccination, as evidenced by the simple positive and negative distribution graph. In this data project, the best model with the highest AUC value is Ridge Regression. However, when using deep learning to train the model, we only picked a few parameters. The deep learning model might perform much better if more parameters are selected and adjusted.

Reference

[1] Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. " O'Reilly Media, Inc.".