Analysis of Reaction to COVID19 Vaccine on Tweet

Liding Li
lidingli@umich.edu
Data Science

Xin Hong
xinhong@umich.edu
Computer Science

1 INTRODUCTION

COVID-19 is a respiratory infectious disease caused by a newly discovered strain of coronavirus. Since the Covid-19 epidemic, there has been quite a buzz in social media platforms and news sites regarding the need for COVID-19 Vaccine. Even though most people agree that vaccines are one of the most effective factors in ending this epidemic, for various reasons, not everyone has the same positive attitude towards COVID-19 vaccines. Understanding the views and attitudes of people on the new coronavirus vaccine and the reasons behind them will help scientists to promote the vaccine in a more targeted manner.

Hence, the goal of our project is to analyze the attitudes of people in different regions towards the COVID-19 vaccine by processing the Twitter tweets with the hashtag #CovidVaccine. We cleaned the data first to remain the useful part and used a LSTM model to provide a sentiment label for each tweet. Then we would build a different neural network model and a decision tree to analyze how time and location influence people's attitude towards the vaccine respectively. Finally, we performed a community detection on the data. Neural networks, decision trees and community detection have not been implemented in previous projects and they are both taught in the second half of the semester.

2 DATA

The data we used is an ongoing collection of tweets with hashtag #CovidVaccine from January 8, 2020, to March 11, 2021, available on the Kaggle platform. There are approximately 170k records, each of which contains a tweet username, a user-defined location, a user description , the date and time of the user account, the number of followers, the number of friends, the number of favorites, date and time of the tweet, and the actual UTF-8 text of the tweet. We mainly used the information from user-defined location, user description, the creation dates of the tweets, and the actual text of the tweets for analysis.

However, not every record contains complete information. For example, some users might not leave their description or location in their profile. Besides, not all the information is in English plaintext form, some of them include emojis, and some of them might be in other languages. Therefore, it was necessary to process the data before we did any further analysis.

3 DATA ANALYSIS

3.1 Q1: How can we attach sentiment label to each tweet?

3.1.1 Data. We will use the 170k original tweets in the data as the basis for sentiment analysis. To ensure that the text can be correctly recognized and reduce misjudgments, we have removed text punctuation and https links, translated non-English texts, converted emoticons into corresponding texts, and removed some custom text styles.

3.1.2 Technique & Challenges. To study people's attitudes towards the new crown vaccine, as well as the influencing factors, we need to know what kind of sentiment each tweet mainly expresses, and we must be able to label each tweet.

Since the outbreak of the COVID-19 epidemic was in 2020 and the clinical trials of most of the COVID vaccines on the market had not started until late 2020, the number of tweets with #CovidVaccine hashtag is not particularly large, which might impose a negative effect on our semantic prediction results. Meanwhile, there are no semantic analysis tools specific to documents related to COVID-19 topics and we were not able to know the real semantic label of each of the collected tweets, so it was difficult to build and train a neural network model from scratch for semantic prediction. Therefore, we decided to use the pre-trained model for labeling.

3.1.3 Experimental Setup. As state above that none of those semantic analysis models is specific to the COVID-19 corpus, we removed the words "covid", "virus", "pandemic" and "vaccine" from each tweet because the word "covid" cannot be recognized by most corpora and although the words "pandemic" and "vaccine" typically have negative meanings under normal circumstances, they are much more neutral in the discussion about the COVID-19. We think this additional processing can make our final emotional judgment more objective and accurate.

We have tried three different pre-trained models to label the emotion of the tweets. We first tried to do the labeling with the Flair model which is based on the LSTM method [1]. It determined whether each tweet conveyed negative or positive sentiment and output a score from negative one to positive one. It took a relatively short time to make the prediction. However, the training data of Flair comes from IMDB data, so it might be incompatible with texts from tweets. We would assume that the label generated by Flair might have a larger error. Then we tried to use Textlob for labeling and got polarity and subjectivity scores for each tweet. However, subjectivity was actually not important for the analysis of our problem, and moreover, Textblob only recognizes emotional keywords and average their polarity scores as the final result. Since most of the tweets were relatively short in our data, Textblob might miss much information. Finally, we used a Recurrent Neural Network model specifically for twitter emotion recognition. It calculates the probabilities for each tweet to belong to each of Ekman's six basic emotions including joy, anger, sadness, surprise, fear, and disgust and labels it with the emotion with the highest probability [2]. This model provided us with more intuitive and diversified tags, so we mainly used the labeling results from this model. However, the first two label sets can still act as references for our subsequent quantitative analysis.

3.2 Q2: How does the emotion change over time?

3.2.1 Data. The data here we used is the same as previous, but we did more preprocessing to do the analysis. We aggregated the data by each date and counted the numbers of joy, fear, and total emotions, and then calculated the ratio of joy and fear. We also calculated the moving average for those two emotions. As you might notice, there is a slight upward in joy percentage and slight downward in fear percentage (Figure 1). The reason we did not choose other labels is that "suprise" is tend to be neutral in sentiment and "sadness", "anger" and "disgust" are rare in our dataset. Further, as a sanity check, we used the labels predicted from the flair model to visualize the trend in ratio. We noticed similar results are delivered (Figure 1). However, given the daily changes from flair model are more flucturate, we will continue to use the emotion model.

3.2.2 Technique & Challenges. Then, we would like to use one of the time series techniques, LSTM, to model the general trends and predict the next ten days' emotion trends. However, the technical challenge was that after aggregating data by date, we have a limited amount of observations (143 dates). It will be very likely to commit to overfitting if the model complexity is high.Therefore, we would use techniques such as cross validation and early stopping to prevent overfitting.

3.2.3 Experimental Setup and Outcome. First, we split the first 100 days as training data, and the rest 43 days as test data. We then built a Neural Network model with ten LSTM units and one dense layer, and used 10 percent validation data for each training epoch to avoid overfitting. Consequenntly, we found the rooted mean square error are 0.04 (for joy) and 0.03 (for fear) on the test sets. As you can see in the graph, the model capatures an upward trend in percentage of joy and a downward trend in percentage of fear (Figure 2). For the future 10 days, the model provides the following results, which align our assumption that there is no huge rise or drop but stays similar as the past (Figure 3).

Therefore, we can conclude that the following ten days will follow the same pattern, and in the long run, rise in the "joy" and down in the "fear" can be observed.

3.3 Q3: How do the location, number of friends and number of followers affect their attitude towards the vaccine?

3.3.1 Data. In this part, we wanted to know whether users' location, number of friends, and number of followers will affect their attitudes towards the Covid vaccine.

We can directly get the numbers of users' friends and followers from the original dataset and get the emotion labels for each tweet from the previous session. However, not all the users show their location in their profiles, and there is no standard format or restriction for location input. Therefore, we first use Pygrapy to analyze users' countries from the user location session or found the countries mentioned in their description session if their user location session were blank. Finally, we got around 89000 tweet records with all location, friends, followers, and emotion labels information.

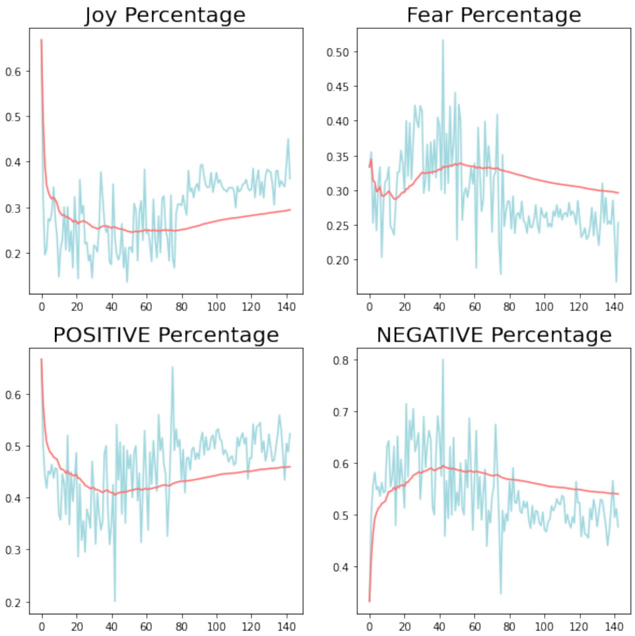

Figure 1: time

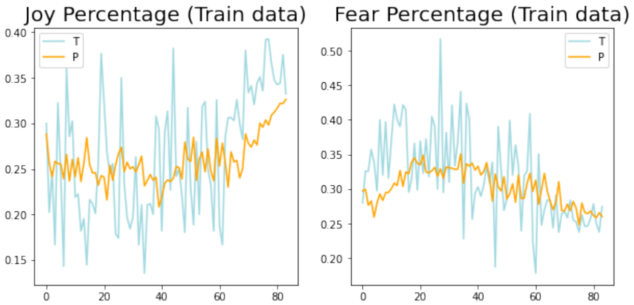

Figure 2: time

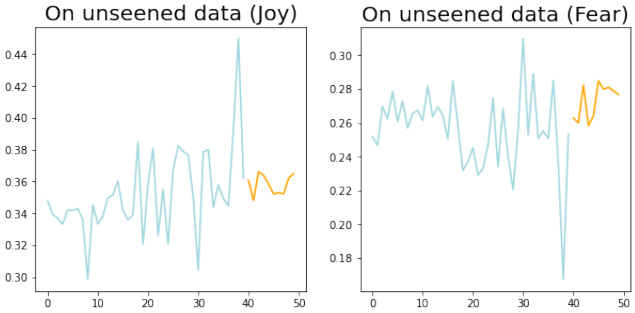

Figure 3: time

3.3.2 Technique & Challenges.

We decided to build a decision tree based on location, number of friends, and number of followers. If the label prediction accuracy for the decision tree could be fairly improved from random guess, then we could infer that these factors have a certain impact on the user's attitude towards vaccines. And we could also see how much each factor contributes to the users' attitude through the model.

The first challenge we encountered was the need for the training data to be in the same type for applying the decision tree in Sklearn. Since location is a categorical variable, we convert it into multiple dummy variables using Pandas. We found 107 countries from the data so this conversion greatly increased the size of each record.

Besides, the original data, as well as the subset of data we used were biased. For example, joy, surprise, and fear dominated the user's emotions we extracted, and most of the users are from America, Britain, and India, and most of the rest are from Canada, South Africa, Ireland, and Australia. Consequently, our model might provide more accurate predictions for tweets from these countries or with these sentiments. To improve the objectivity of our model, set the class weight for classifiers as balanced. We also tried random forest and used cross validation.

3.3.3 Experimental Setup and Observations. We simultaneously applied the decision tree and random forest on the data with cross-validation. We found that the overall accuracy of label prediction was around 38%, which was slightly larger than twice the accuracy for the random guess. It might indicate that these factors do have an influence on users' attitude towards the Covid vaccine. The feature importance of the followers amount is 46.6% and that of the number of friends amount is 45.8% while the sum of feature importance of all the countries is only 8%, which implies that location might have little impact on users' emotion.

Considered that we only have several records from users for some countries, we then tried eliminating those records to train the model. We also tried only training on records from those dominated countries and records with those dominated emotions to make our model more general, but we got very similar prediction accuracies

Then we tried discarding the location factor, only considering the follower amount and the friend amount. The accuracies we got from the decision tree and random forest could be up to around 60%, indicating that those factors are fairly influential for users' attitude towards the vaccine. This result is understandable to an extent. Since the follower amount and friend amount can reflect how users are engaged in Twitter or other social media. People with different levels of engagement with online social media might have different levels of real-time mastery and sensitivity to information, and they might see different aspects of reports about the Covid vaccine, so they are likely to have different attitudes towards the vaccines.

However, we still believe that location might have a certain relationship with users' emotions on the vaccine. We first tried to apply greedy forward selection to important features among those countries but we failed to get candidate variables. Then, we applied the random forest algorithm to the three kinds of data and extract the top ten important features except for friends amount and followers amount. Then we plot the KDE curves for those selected countries including to see the distribution of the emotions in each country. We found that the peaks occurred on joy and fear and the densities of joy and fear are similar, but the density of joy in Bangladesh and Uruguay is much larger than that of fear compared to the statistics in other slected countries.

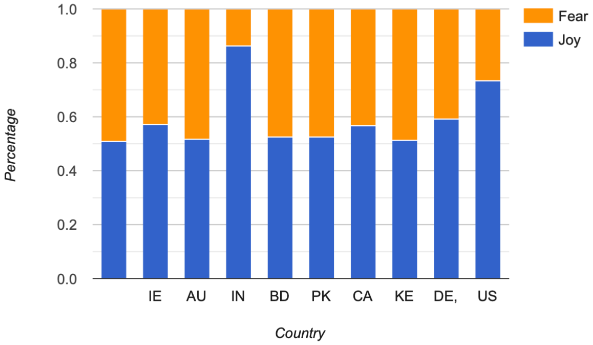

Figure 4: country difference

3.4 Q4: Is there any pattern in the community group?

3.4.1 Data. In this session, we want to detect the community from the collected tweets to see whether there is any pattern of emotions of users in the same community. Therefore, we used the user names in the original data, the user names mentioned in the tweet text, and the emotion labels of each tweet we extracted in the previous session as the data for our analysis.

3.4.2 Technique & Challenges. We matched the users to the users they mentioned in the tweets, and by default, we considered users share the same emotions about the covid vaccine with the users they mentioned. We regarded all users as nodes and established edges with a weight of 1 between the users and the users they mentioned.

Due to the limitations of computing power and resources, we could only use sample data rather than all the data to build the graph. However, as mentioned earlier, since this is a relatively new data set, its volume is not very large and it is loose to an extent, which means that few users mentioned each other in this data set. Therefore, the challenge we face was that the low density of the established graph. In order to connect more users, we added edges among users who were mentioned by the same users with a weight of 0.4 and added edges among users who mentioned the same users with a weight of 0.8.

3.4.3 Experimental Setup. We tried to use the Girvan-Newman's method for community detection. We repeatedly calculated the betweenness centrality of all edges in the graph and removed the edge with the highest betweenness score until there are no edges remain. However, this method did not give a good performance because the low density of the graph prevented it for giving a more interpretable result.

Then we switched to implementing the spectral clustering, computing the Laplacian matrix, determining the eigenvalues and eigenvectors, and then applying K-means to do the clustering. During

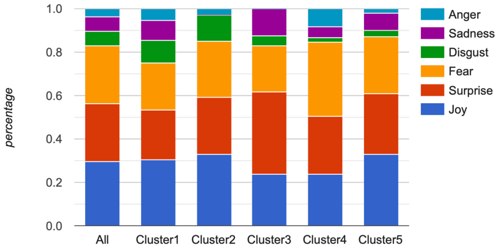

Figure 5: cluster percentage

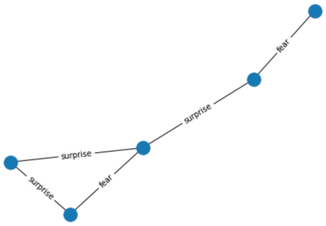

Figure 6: cluster percentage

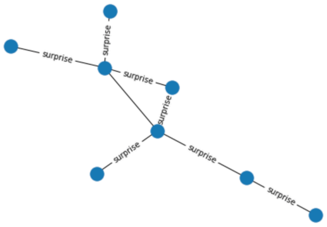

Figure 7: cluster percentage

this process, we used the Elbow method to determine the number of clusters.

3.4.4 Observations. We wanted to know whether there were some emotion patterns among users in the same community. Intuitively, we computed the percentage of emotion in each community. In the sample data, joy accounted for 29.6%, surprise accounted for 28.8%, fear accounted for 28.7%, sadness accounted for 6.76%, anger accounted for 3.7%, and disgust accounted for 2.36%. The distributions of the emotions in most of the groups were similar to the sample data. The figure 5 shows the emotion distribution of several clusters. It makes sense because there are not many highly connected subgraphs.

However, we still found some noteworthy clusters where users share similar emotions as we expected. Figure 6 and Figure 6 are two examples.

4 CONCLUSIONS & DISCUSSION

4.0.1 Key observations from the analysis. Overall through the three parts of analysis, we have found these information. 1) Positive emotion will continue to increase and negative emotion will decrease, as the vaccine is spread and is proved to be effective.

2) We have found numbers of followers and friends place greater effect on user emotion than locations. Though there are some countries having very unbalanced positive and negative emotions, most of the counties have similar amounts of negatives and positives.

3) We obesevered that the clusters have a similar distribution of emotions as the sampled dataset. But we did detect some patterns of emotion that some users share with others.

4.0.2 Challenges. The biggest challenge is that our data is quite messy. Deciding which words to keep and remove will directly affect the result of analysis. For instance, if we keep pandemic related keywords, more negative tweets will appear, since our pre-trained model was not trained on COVID19 related text.

The second challenge is that we have a limited amount of days to analyze the trend, and thus fitting a Neural Network model will very likely to overfit the data. So we reduced the complexity of the model as much as possible and used some techniques to prevent overfitting.

The third challenge is that we found the data is unbalanced when performing location and feature analysis. Most of the labels are joy and fear and a few of them are other labels. Further, converting countries to binary features brings a sparse matrix, so we have to select some out of all features in order to perform analysis.

The last challenge is that we have a sparse but huge adjacency matrix when building graphs. Therefore, we decided to choose only 3000 of users in the end.

4.0.3 Things learned. There are many components of the projects, so we think we learned many things. For example, we learned how to build neural networks to predict time series data, how to analyze features after fitting a classification model, and how to build graphs and perform community detection.

4.0.4 Favorite parts of the project. The favorite parts are time series modeling and community detection. For the time series modeling, we need to build the model using keras and tuning each parameter. For community detection, we need to write the algorithm our own. Hence both parts give us sufficient opportunities to know the model and the algorithm.

4.0.5 Team members' contribution. Liding finished the label extraction and time series analysis. Xin finished feature analysis. We finished community detection and wrote the report together.

5 REFERENCE

[1] Twitter Emotion Recognition, Github link:https://github.com/nikicc/twitter-emotion-recognition

[2] Flair NLP Models, Github link:https://github.com/flairNLP/flair