What is People's Emotion Under COVID-19 Pandemic

Yuhang Zhou
tonyzhou@umich.edu
Data Science

Jing Zhu
jingzhuu@umich.edu
Computer Science

1 INTRODUCTION

During the past three months, the COVID-19 pandemic outbreak has influenced nearly all the corners in the world. Policies like "travel ban" , "social distancing" or "stay-at-home order" have caused deep controversy and hot discussion in the social media all over the world. Twitter is a popular microblogging platform in the US. People's tweets with respect to COVID-19 provides novel insights in understanding how people feels about COVID-19.

Our project focuses on understanding people's emotions with respect to COVID-19 during this special period of time. In our project, we collect the tweets associated with the COVID-19 from January 21 to March 12. We first apply the recurrent neural network to extract the emotion in the tweets. Then we use the decision tree with user information to predict the emotion of the tweets. Decision tree is taught after the mid-term and we never implemented it in the previous projects. In the meanwhile, we also applied the spectral clustering to detect the communities, which is also not implemented in the course projects.

2 DATA

The data we use for the analysis is a recently released twitter dataset associated with the novel coronavirus COVID-19 (SARS-CoV-2) [2]. The dataset is the ongoing collection of tweet IDs about COVID-19 from January 21 to April 10. At the time of data preparation, the dataset contains the tweet IDs from January 21 to March 12, and for our data analysis project, we use this part of data to do the analysis. Also, note that the data for Feb.23 miss in the original dataset. And so it is also missed in our dataset.

Using the tweet IDs they provided, we apply the twitter API to hydrate the text of the tweet along with other information. The tweets hydrated contains a variety of information such as created time, ids, or hashtags. We have uploaded part of the datasets hydrated from twitter in Google Drive[1]. We collect 40 million tweets and relevant information in total.

3 DATA ANALYSIS

3.1 Q1: How to predict emotion from tweets

Since our main focus is on understanding how people feel about COVID-19. The first step that we need to do is to find some ways to classify the emotions from the twitter contexts.

3.1.1 Data. We use the full text contents of tweets and classify each tweet into one emotion.

3.1.2 Techniques and Challenges. When we use the recurrent neural network (rnn) model for predicting emotions from English tweets. Due to the fact that training a rnn to predict emotion requires huge amount of data and label and lots of computing resources, and COVID-19 dataset is a new dataset without ground truth labels of emotions, we used the pretrained character-based recurrent neural network model from [3].

We meet a couple of challenges in this part. First, the model and the code they provide is too old and it does not work. We have to rewrite the code to make it compatible with the theano environment that we are able to install. Second, our dataset is pretty large and the prediction is slow. It takes 24 hours to predict emotion for the data in January. And we address the problem in the following way. First, we revise the code into multithread, hoping that this will make the prediction faster. But we found that the theano package that we are using will slow down the process because theano does not work until there is enough room in the cache. And this greatly slows down our multithread process. Finally, we split the data into 20 parts and launch 20 emotion prediction jobs at the same time. It fastens greatly our prediction. It takes us less than 6 hours to complete all of the predictions using this approach.

3.1.3 Experimental Setup. We translate the emotion recognition problem to an emotion classification problem by using Ekman's six basic emotions. Paul Ekman studied facial expressions to define a set of six universally recognizable basic emotions: anger, disgust, fear, joy, sadness and surprise [3].Ekman's six basic emotions are one of the most popular emotion classifications that researchers use today. [3].

Besides, the model is able to predict emotions in two settings: multiclass and multilabel. In multiclass setting, the most obvious emotional hashtag is set as the target and for each text, there is only one target emotion. In multilabel setting, the classfifer will give a prediction for each emotion category of whether this emotion is expressed in the text. In order to simply our further analysis, we use the multiclass setting. For each twitter text, it will just predict the most obvious emotion from that context.

3.1.4 Observations. We are unable to give a quantitative result on how good the prediction is because we do not have ground truth label for the dataset. But from our observation, we think that most predictions the model gives are pretty accurate. Table 1 shows some of the accurate contexts and the predictions that our model made. However, as shown in Table 2, there still exist some cases that the model can not classify perfectly.

We also want to see what kind of emotion dominates for COVID-19. Table 3 shows the fraction of each emotions for each month. It is clearly indicated that fear is the most dominated emotion for COVID-19.

[1]For details of the dataset, check https://drive.google.com/drive/folders/1V92l5sCMnv-pcOypWkCCfHho620_e075

Twitter Context	Predicted Emotion
Wait. So you all are telling me that there is a corona virus/plague outbreak and now locust swarms? Does this sound fami?	Fear
Trump Says That 'Warmer Weather' Will Make Coronavirus Disappear and Everyone Thinks He's an Idiot	Anger
This is so sad	Sadness
The coronavirus outbreak could derail Xi Jinping's dreams of a Chinese century	Fear
This is sad, Dr. Li, the doctor who warned the world about CoronaVirus has died. RIP Dr. LI, you are a true Hero!	Sadness
Love this long synthetic jacket. Took 3 months to arrive from China. Luckily arrived ahead of the Coronavirus.	Joy
a pakistani student in china, a friend of mine, sharing the problems they are facing out there, they are so scared	Fear

Table 1: Qualitative examples for emotion prediction

Twitter Context	Predicted Emotion
Actually, no western nations will take reference from WHO (Department of health in China) advice	Surprise
realDonaldTrump: Just had a long and very good conversation by phone with President Xi of China. He is strong, sharp...	Joy
Your health should never be determined by who you are, what you look like, or whom you love	Joy

Table 2: Failure Cases

	Fear	Joy	Surprise	Sadness	Disgust	Anger
Jan	57.23%	28.05%	10.94%	2.31%	0.75%	0.71%
Feb	49.33 %	29.62%	16.14%	2.99%	1.05%	0.86%
Mar	51.38%	29.28%	14.03%	3.28%	1.06%	0.96%

Table 3: Fraction of each emotion in three months

We also track the number of fear tweets through time, presented in Figure 1. The trend also coincides with the time of COVID-19 outbreak. In late January, there are quite a lot tweets expressing that they feel fear about the COVID-19 outbreak. It was the time that China was experiencing a significant growth of COVID-19 patient numbers. But after that, the number of fear tweets drops. Throughout February, the number of fear tweets remains almost a constant, indicating that most people do not feel so scared about COVID-19. But it increases tremendously in the last two days of February and early March. This is exactly the time that community spread of COVID-19 was reported in US and Europe and number of cases of COVID-19 in US start to increase rapidly. Since the trend in Figure 1 follows our expectation, we can also say that the prediction given by the model is accurate in general.

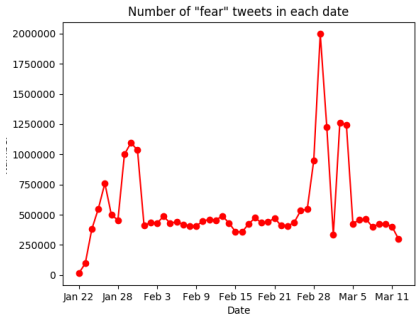

Figure 1: Number of fear tweets from Jan 22 to March 12

3.2 Q2: How does the user information influence emotions

In this part, we want to figure out how user information: time and location influences their feelings about COVID-19. Is there a particular time period or people in some places are more likely to get scared for COVID-19?

3.2.1 Data. Since for this part, we want to analyze how user information influences their feelings about COVID-19. We use the time, location and emotion labels that we get from 3.1.

3.2.2 Technique & Challenges. We want to figure out if there is any specific user information that highly influences the people's emotion of COVID-19. We choose to use decision tree because it's a highly interpretable model and it's taught after mid-term and not implemented before.

We met two main challenges for this part. First, our dataset contains 40 million tweets and it's computationally hard for us to train a decision tree on such a large dataset. We do not have so many computing resources. Our solution to this challenge is that we will random sample 10000 data from the dataset and use 7,000 for training and 3,000 for testing. Second, the meta data of tweets contains the exact seconds when user creates the tweet. It's hard to find tweets that coincides with the other in seconds. So in order to learn a more interpretable feature, we truncate the time data and just keep the days of the tweets.

3.2.3 Experimental Setup. We first random sample 10000 tweets from the dataset. Then we keep only the time, location and emotion of the original tweets. And for the time of each tweet, we only keep the date that the tweet was created. We split the dataset into two parts, 7000 tweets for training the decision tree model and 3000 for testing its accuracy. Then we encode the features using one hot encoding method and apply the decision tree implemented in sklearn to fit our model. It is important to do one hot encoding for features because the decision tree in sklearn package does not handle categorical data. If you do label encoding and convert the features into integers, then the integers will have relative order. But

our original data, both date and location are nominal data that does not have relative order. The decision tree will output one of the six emotions.

3.2.4 Observations. The accuracy for the decision is 48.9% while the accuracy for random guess is 1/6 = 16.7%. We can thus infer that the time and location actually has a great impact on people's feelings for COVID-19. we also found that the 5 most influential time and locations features for emotions and these are listed in Table 4 along with their importance factor. From this table, we can observe that the top five most influential factors are also time features and also the turning points in Figure 1.

Feature	Importance
Jan 30	0.0114
Mar 03	0.0086
Mar 04	0.0085
Feb 28	0.0070
Mar 06	0.0059

Table 4: Top 5 most influential features for decision Tree

Besides, the most important location feature for the decision tree is united states with importance 0.0039. It implies that our decision tree model captures the important time and location features.

3.3 Q3: How does emotion spread in the communities

In this part, we want to detect communities through all the tweets we have. Besides, we want to find out if there is any common emotion that can spread in the communities.

3.3.1 Data. We use the mentions, name of the user, and the full text of the tweets from our dataset to do community detection.

3.3.2 Technique & Challenges. We use spectral clustering to do community detection and identify the emotion that spread among them. We construct the graph with the mentions in the tweets as edges. Since the graph is not fully connected, we extract the connected subgraphs and do the spectral clustering to detect the communities on these subgraphs.

The main challenges that we encountered in this part are as follows. First, we have too much data and it's computationally hard for us to construct graphs using all of them. So we choose first 100,000 of tweets among them and construct graphs based on the 100,000 tweets. Second, we want to quantitative evaluate if people in one cluster shares the same emotion. Due to the limit of time and computation resources, we switch to give qualitative examples on how emotions spread in communities.

3.3.3 Experimental Setup. The method of how to construct and visualize graphs from tweets and spectral clustering is brought partially from [1]. If in the tweets, user1 mentions user2, then there is an edge between user1 and user2. It also means if user1 and user2 are connected by an edge, both of them share one tweet. We classify the emotion of this tweet with the model mentioned in Section 3.1. An edge represents a tweet and also represents an emotion. From all the graphs that we construct from the 100,000 tweets, some subgraphs are chosen for spectral clustering.

For spectral clustering, we first calculated the graph Laplacian L, obtain the eigenvalues and eigenvectors L and then do clustering based on the order of the eigenvalues. For each subgraph, the number of clusters is determined by the silhouette scores.

3.3.4 Observations. Figure 2 and figure 3 are two examples of the spectral clustering in the subgraphs. We choose to cluster the subgraphs into two communities due to the subgraph order and silhouette score. These figures show that the spectral clustering successfully detect two "well-separated" communities, which follows our expectation. We also label the emotion on the edge, and it suggests that both "yellow" and "purple" communities share the fear emotion.

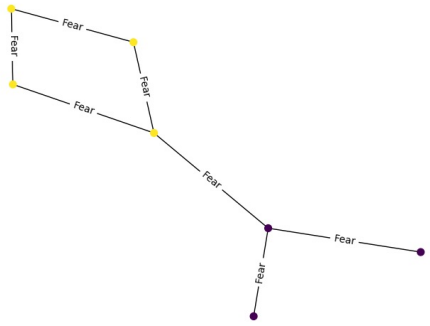

Figure 2: Unweighted cluster of seven people

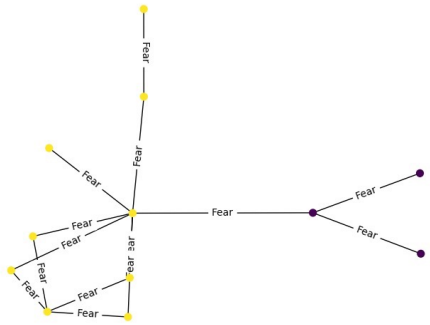

Figure 3: Unweighted cluster of twelve people

We can conclude that our spectral clustering method are effective to detect the communities in our dataset and in the communities, it may have more probability to share the same emotion, especially for fear emotion. Later, we may do some quantitative evaluation to demonstrate the implication.

4 CONCLUSIONS & DISCUSSION

4.1 Key Observation

We find three key observations from our analysis. First, fear is the most common emotion that people have for COVID-19. Second,

people's emotions towards COVID-19 are highly influenced by time. The late January and early March is the most influential time period for people's emotions of COVID-19. Late January is the time that COVID-19 has an outbreak in China and in early March, the US is experiencing a COVID-19. Third, fear is a shared emotion in twitter communities and fear may also spread between communities.

4.2 Challenges

We face two main challenges for this project. The first is that the size of our dataset is too big for many of our analyzing methods. We face serious time and memory issue. So we have to apply multithread and random sample technique for our dataset. In the meanwhile, we believe that this is an issue for quite a lot data mining problems today.

The second main challenge that we are facing is model selection and model interpretation. We seek to provide some interesting insights for COVID-19 and so we need our model to have high interpretability instead of just having a high accuracy. We spent a lot of time on finding out the most influential features in decision trees and visualize the communities and their emotions.

4.3 Take aways from this project

For this project, we hydrate the tweets from twitter ourselves and do feature selection ourselves. It makes us to first think about the questions that I want to answer and what data we need to answer this question instead of blinding applying data mining techniques.

Second, in the class when we are talking about various techniques like BFR and Hadoop, Danai always emphasize a lot on the time and memory issue. The final project gives me a sense of how big these datasets are and why it is so important to solve the time and memory issues.

Third, we feel that interpretability is important for data mining problem. For example, when we do the community detection part, first we only draw out the communities given by spectral clustering. But no one can know anything about COVID-19 merely through the communities. The more important thing is if there is any shared features in the communities and if it can provide some interesting insights.

4.4 Favorite Part of the project

Our favorite part of the project is the community detection part and understanding how emotions spread in communities. We learned how to construct a graph from the raw twitter data. We had a hands-on experience on using spectral clustering to do community detection. And We found an interesting insight through community detection. We found that fear is actually an emotion that can spread in communities! It's a brand-new insight.

4.5 Contribution

Both of us contributed equally for this project. We brainstormed the idea of this project together. We hydrate tweets from twitter at the same time. For the coding part, Yuhang Zhou writes the data preprocessing and feature extraction, the emotion prediction, the decision tree and graph construction and the cluster visualization parts. Jing Zhu works on changing the emotion prediction to multithread, training the decision tree and outputting the meaninful features and the draft of spectral clustering. As for the writing of the report, Jing Zhu focuses on writing the report and Yuhang Zhou focuses on proofreading the report.

REFERENCES

[1] Twitter community detection. https://github.com/benosment/twitter-community-detection. Accessed: 2020-04-19.

[2] Emily Chen, Kristina Lerman, and Emilio Ferrara. Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*, 2020.

[3] Niko Colneriĉ and Janez Demsar. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE transactions on affective computing*, 2018.