Leveraging Twitter Data and Deep Learning Method for COVID-19 Case Prediction in United States

Shichen Li†
School of Mathematics
Georgia Institute of Technology
Atlanta, Georgia
sli828@gatech.edu

Siyan Cai
The H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, Georgia
scai70@gatech.edu

Tongzhou Yu
College of Computing
Georgia Institute of Technology
Atlanta, Georgia
tyu310@gatech.edu

1.INTRODUCTION

Coronavirus disease or COVID-19 is a novel virus that started at the end of 2019. It is a new infectious disease spreading through respiratory droplets and contact and is generally infectious to human beings [5]. Since its appearance, COVID-19 has quickly spread around the world, with over 200,000,000 confirmed cases and 4,000,000 reported deaths in more than 200 countries at present time [1]. Thus, predicting the spread and development of COVID-19 has become a worldwide hot topic. However, accurate data collection can be hard and untimely during a serious pandemic, while overloading real-time feedback directly generated by the public is all over the internet. Specifically, social media platforms are the top source for such information and can be leveraged to analyze different aspects of the ongoing outbreak. Twitter is one of the most active social platforms in the United States where people may share information regarding the pandemic. Plenty of research has been done using Natural Language Processing (NLP) techniques such as term frequency, sentiment analysis, topic modeling to study the public opinion on different issues. Given the current pandemic, many twitter datasets have been created by collecting tweets including COVID-19 related keywords. Based on these datasets, many NLP techniques could be employed to extract useful features for understanding the development of COVID-19 and perhaps predicting its future spread.

2.RESPONSE TO COMMENTS

Insufficiency of only using tweets

We plan to compare tweet feature with other features such as google search to find out the importance of tweets' information. Details are included in Section 4.3. For now, we have finished using only tweet feature for prediction.

Insufficiency of only predicting covid cases

Instead of only predicting cases, we predict both covid cases and death.

Extensions including case studies and hypothesis testing

For now, we have finished the covid prediction in California. We plan to carry out similar prediction for other states. As for hypothesis tests, we observed the scatterplots for positive sentiment and negative sentiment vs daily new confirmed cases, and found no obvious patterns indicating linear relationships, so we've decided not to carry out hypothesis tests on these two features. We will investigate other features in the future.

More Reference

We added references [11] and [13].

3.PROBLEM DEFINITION

During the current COVID-19 pandemic, given the daily infection statistical data from JHU and available public opinions from Twitter, we aim to accurately forecast future development of the disease in the state of Georgia. We will focus on extracting useful insights from tweets and finding the optimal prediction model.

4.RELATED WORK AND SURVEY

4.1 Understand tweets with NLP

Natural Language Processing techniques are essential for our study, for they transform the textual data of millions of tweets into explainable features like fear and anxiety, which will be of great use in predicting the pandemic. There are millions of tweets every day even if we only consider those related to COVID-19, therefore, data pre-processing is very important. Researchers performed comprehensive processing steps to clean the data, which includes but is not limited to filtering repeated tweets, URLs, symbols and stemming the tokens of tweets [2][3]. Given a feature such as the emotion "fear", there are plenty of classification techniques we can choose from during our NLP process, such as Decision Tree, Support Vector Machine, Random Forest, Logistic Regression and KNN models. In all three research, various classification techniques are tested and compared upon given features [1][2][3]. Among them, Naïve Bayes and Decision Tree outperform others, with better accuracy [3][4]. Some research didn't pay much attention to feature extraction but laid great importance in detailed analysis upon tweets and features themselves [4]. They found that these methods prefer shorter tweets, and will have a better performance when

using KL, Euclidean distance. However, to better apply NLP, some researchers also several tested classification techniques under different feature extraction methods [2]. The interesting result is that although Naïve Bayes and Decision Tree are great in some cases, their performances vary greatly when work with various feature extraction methods, like bigram and trigram [2][3]. Logistic Regression and KNN methods, which don't perform the best in many cases, surprisingly show a consistent performance in all cases, but luckily, we found that Decision Tree model still performs the best in most cases [2]. However, some technical details remain untested in all mentioned research, such as optimization method which might greatly influence the performance of some classification methods mentioned above.

4.2 Choose proper tweet features

Next, it's important to study how other researchers applied different NLP techniques to generate useful features to solve their problem. Boon-Itt, et. Al extracted three kinds of features to observe the trend of public awareness and perception of COVID-19, including keywords, sentiment, and common topic/theme [5]. This study points to us the possible methods of features extraction for our purpose. Also, researchers find that even simple feature such as the google searches for 'wash hands' could also help predict the speed of the spread of COVID-19 [6]. Furthermore, we could use other keywords on the tweet dataset, evaluate their performance and then find most useful keywords for prediction. In Deb, et. al's work, features from more than one platform are used for analyzing airline stock price volatility during COVID-19 pandemic, which also gives us some insights. Deb performs frequent term searches on relevant twitter data from selected accounts to obtain key topics of interest and used Google Trends to compute daily Google Search volume for each topic as predictors [7]. Such way of combining data from both twitter and google lift restrictions of using only one source dataset when extracting features.

4.3 Other features

Other features could also be useful for covid prediction. Recent study compares twitter data and Google Trend data and analysis their performance on detection of COVID-19 waves in US and Canada [13]. They find that in some period twitter data has better performance for prediction while in some other period Google trend has better performance. Also, in different states or countries, twitter and Google Trend might have different performance. In our work, we plan to also start with comparing twitter and Google Trend data and then add other features.

4.4 Predict COVID_19 with LSTM

Apart from researching areas in Natural Language Processing, it's crucial to examine works on COVID-19 case prediction. Omran, et. al's comparative study between LSTM and GRU shows how to leverage state-of-the-art deep learning models on predicting time-series data [8]. Their conclusion suggests us to use LSTM for confirmed cases forecast and to rely on RMSE, MAE, MAPE for evaluation, which is helpful in forming the methodology for the predictive part of the project. However, this study only utilized historical COVID data, while Bhimala, et. al's approach on COVID-19 infection prediction from weather data provides the most valuable reference [9]. The authors focused on different regions of India and studied the correlations with meteorological features. Four predictors (humidity, maximum temperature, minimum temperature, mean temperature) are each used once in combination with previous COVID-19 cases data to train multivariate LSTM (Long Short-Term Memory) deep learning model for forecasting. Results showed that univariate LSTM with only previous cases was the most accurate for short-term prediction, while the other weather factors helped improve long-term prediction. This paper presents a workable approach applicable to our project, and we would be able to generate useful predictors based on other NLP references.

5.DATA COLLECTION

5.1 Data Collection Process

COVID-19 Tweet Data

Our project will use the COVID-19 Twitter chatter dataset provided by USC [10]. This is a data set that has filtered out all COVID-19 related tweets and stored their tweet id in terms of day and hour.

For each tweet, we first use twitter's official API to get the text of it, which could be useful by performing different NLP techniques on it to extract valuable features that could be later used to predict the development of COVID-19.

In order to evaluate the performance in different locations, the location of each tweet is required. Since location is not a necessity when posting a tweet, over 99.5% of tweets lack location information. In similar work which also analyzes COVID-19 tweet data, they use self-reported user profile locations instead [11]. We implement a fuzzy text matching algorithm that matches profile location with state name or their abbreviation codes inside United States. We find that around 60% of the users' self-reported profile location are legitimate and around 18% of them can be matched with a state in the United States.

We intend to use four months of tweet data from 2020-08-01 to 2020-11-30, among which the first 3 months are used to train and the last month is used to test and evaluate. (Considering the volume of the dataset and the limit speed of Twitter API, we will randomly sample 1,500 tweets inside United States each hour.)

COVID-19 Case Data

We will use the COVID-19 dataset provide by John Hopkins University [12]. This dataset has the number of newly infected, dead, recovered people in each state of the United States and in each country around the world every day. Corresponding to our twitter dataset, we will use the infection, mortality data from 2020-08-01 to 2020-11-30 for.

5.2 Data prepressing

Given the tweets data, we employed several measures to clean the text before moving on to feature engineering. Upon examining some data, we found emojis used in a large proportion of the sample. To capture the information contained in emojis, we applied the "demojize" function from the "emoji" package to

replace the icon with its meaning, such as smile, angry. Next, we noticed that some tweets were cut off in the middle of a sentence and ended with ellipsis, so we removed these incomplete words at the end of a sentence, as they could not provide any meaning. Additional cleaning included removing the word 'RT', '@userid', hyperlinks, punctuations, numbers, and other special characters. After each tweet has been cleaned, we used NLTK's word tokenizer to split the text into individual words, and then filtered out stopwords using the appropriate dictionary from NLTK. Stopwords are structural words that have little meaning to the analysis. The remaining tokens with length above 1 are first joined back together into a sentence for sentiment analysis and then lemmatized for other NLP techniques. Lemmatization was done with the NLTK package, and it converted each word back to its dictionary form and striped any tense.

6.APPROACH&ALGORITHM

6.1 General Approach

We have primarily found some features to extract. As the project moves on, we might find other useful but previously ignored features that may reveal hidden facts of the spread of COVID-19.

Sentiment analysis feature:

Sentiment analysis classified tweets into positive, negative, and neutral. The proportions of these three categories might reflect public attitude towards COVID and thus function as the feature for COVID prediction.

Frequency of Keywords feature:

The number of tweets including certain keywords might be useful when predicting COVID. Top keywords such as 'death', 'outbreak', or even 'wash hand' are generated first by analyzing tweet dataset.

Emotion analysis feature:

Emotion analysis classified tweets into different emotions, such as fear, happy or sad. Like sentiment analysis feature, these features also reflect public attitudes towards covid and may be used to predict the development of covid.

In the second step, we plan to apply the commonly used NLP model, BERT to do the sentimental analysis. More specifically, we want to apply Decision Tree algorithms which according to literatures are one of the best classification techniques in tweets' emotion detection.

LSTM is chosen as the predictive model based on research. We plan to run a univariate model with pure COVID data as baseline and test out combinations of features extracted from twitter data for comparison. Each model will be trained with three- or five-fold cross validation and compared using the same validation set. The optimal model will be trained again using both the training and validation data and then tested on the testing data for final performance measure.

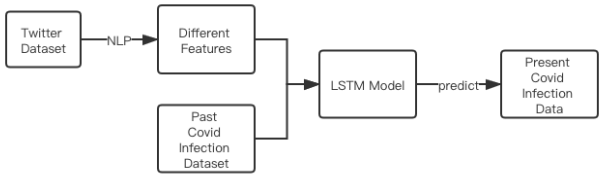

Figure 1 Approach Process Diagram

6.2 Algorithms involved

NLP for Sentimental Analysis

The first set of features is generated from sentiment analysis. From the cleaned tweet sentences, we used the TextBlob package to calculate a polarity score for each tweet. TextBlob will read through all words and return a value within the range of [-1,1], where a negative value indicates negative sentiment and vice vera. We labeled each tweet with "positive", "negative", or "neutral" label, and aggregated the counts for each day to calculate each type of sentiment's percentage compared to all tweets. Figure 2 shows the daily sentiment percentage over the four months.

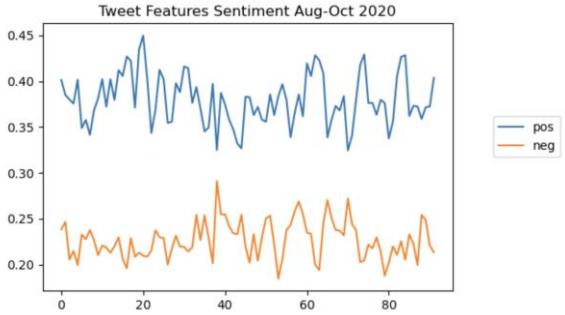

Figure 2 Tweet Features Sentiment Graph

Additionally, we constructed categorical keyword lists and measured daily frequency for each category. Based on research and intuition, our wordlists are summarized in Table 1.

Table 1 Categorical Keyword List

Category	Keywords
Fear	Afraid, fear, worry, nervous, dread, scare, terrify, lockdown
Stay_home	Stayhomechallenge, stayathome, stayhome
Mask	Mask

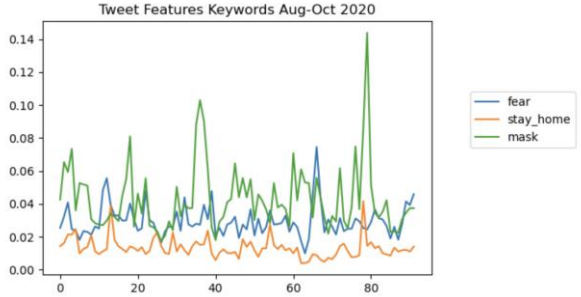

Figure 3 Tweet Features Keywords Graph

LSTM for Time Series Prediction

Time series problems are difficult for they add the complexity of a sequence dependence among the input variables.

The Long Short-Term Memory network (LSTM) is a type of recurrent neural network that keeps track of arbitrary long-term and short-term dependencies in the input sequences. Thus, it is efficient in solving time series problems.

In our project, we set the look-back steps to the latest 3 days so that we can take the most relevant historical development of COIVD-19 into account. These include not only data of confirmed cases but also tweets' sentiment and keywords.

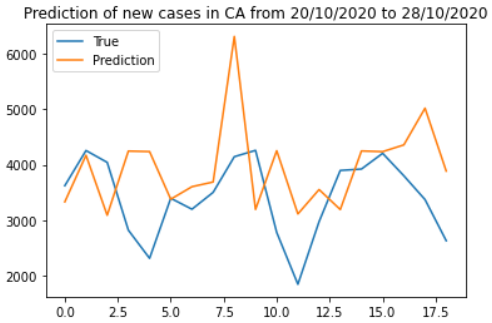

Figure 4 New cases Prediction

As shown in figure 4, our current LSTM model has captured the trend of COVID-19, but showed not much improvement comparing to the base-case model that doesn't take tweets' features into account. Besides, we analyzed the importance of our features, except historical cases' data, negative sentiment and fear features from tweets both ranked top, while keywords like 'stay', 'mask' are about half of the importance as fear.

7.EVALUATION METHOD

As stated in the approach, each LSTM model with different features will be evaluated using the validation set and the final model will be tested with the testing data. Evaluation metrics used for LSTM models are the following: RMSE (Root Mean Squared Error) will be the main objective function to minimize and reflect the prediction accuracy; MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) will also be used to compare the model performance.

8.RESULT ANALYSIS

When predicting daily new confirmed cases, we trained our LSTM model with data from August, 2020 to October, 2020. As a result, we found that the RMSE error of prediction which take tweets' features into account is 15% smaller than that taking only historical COVID-19 confirmed data into account.

9.Difficulties

Data collection:

The number of tweets generated everyday related to COVID-19 is out of our computing capacity, so we could only use a small sample per day and calculate all features in percentages or frequencies. This may cause failure to capture useful information.

Feature Engineering:

It's a very difficult task to generate useful features from raw tweet texts, we've relied mostly on ready-to-use NLP packages and insights from related works research.

LSTM Prediction:

Building a LSTM model is not enough, we still have a lot to improve. Improvements are possible in activation functions, optimization methods and many parts, therefore, improving LSTM is actually much more complicated than simply building it.

10.Uncomplished Part

Case Study & Hypothesis Testing

Case studies and hypothesis testing are left unfinished, which is one of our primary goals in later study. We are going to do a specific case study in California. Besides, we plan to test if increasing tweet sentiment corresponds to an increase in COVID-19 confirmed cases.

Feature Comparison

Till now, we have tested our model in a tweet-only case, however, to discover the importance of tweets in predicting COVID-19 pandemic, features other than tweets should be involved, such as Google search. Thus, in later study, we will study both features and compare their abilities in revealing the true pandemic status.

REFERENCES

[1] Coronavirus Cases:. (n.d.). Retrieved from https://www.worldometers.info/coronavirus/

[2] Zhang, X., Saleh, H., Younis, E. M., Sahal, R., & Ali, A. A. (2020). Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System. Complexity, 2020, 1-10. doi:10.1155/2020/6688912

[3] Samuel, J., Rahman, M. M., Ali, G., Esawi, E., & Samuel, Y. (2020). COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. doi:10.31234/osf.io/sw2dn

[4] Jim Samuel, G. G. Md. Nawaz Ali, Md. Mokhlesur Rahman, Ek Esawi, Yana Samuel. (2020). COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. Arxiv preprint, arXiv:2005.10898.

[5] Boon-Itt, S., & Skunkan, Y. (2020). Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. JMIR Public Health and Surveillance, 6(4). doi:10.2196/21978

[6] Lin, Y., Liu, C., & Chiu, Y. (2020). Google searches for the keywords of "wash hands" predict the speed of national spread of COVID-19 outbreak among 21 countries. Brain, Behavior, and Immunity, 87, 30-32. doi: 10.1016/j.bbi.2020.04.020

[7] Deb, S. (2021). Analyzing airlines stock price volatility during COVID‐19 pandemic through internet search data. International Journal of Finance & Economics. doi:10.1002/ijfe.2490

[8] Omran, N. F., Abd-el Ghany, S. F., Saleh, H., Ali, A. A., Gumaei, A., & Al-Rakhami, M. (2021). Applying Deep Learning Methods on Time-Series Data for Forecasting COVID-19 in Egypt, Kuwait, and Saudi Arabia. Complexity, 2021, 1–13. doi:10.1155/2021/6686745

[9] Bhimala, K. R., Patra, G. K., Mopuri, R., & Mutheneni, S. R. (2021). Prediction of COVID‐19 cases using the weather integrated deep learning approach for India. Transboundary and Emerging Diseases. doi: 10.1111/tbed.14102

[10] Chen E, Lerman K, Ferrara E Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set JMIR Public Health Surveillance 2020;6(2):e19273 DOI: 10.2196/19273 PMID: 32427106

[11] Chen E, Lerman K, Ferrara E Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set JMIR Public Health Surveillance 2020;6(2):e19273 DOI: 10.2196/19273 PMID: 32427106

[12] CSSEGISandData. (n.d.). CSSEGISandData/COVID-19: Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE. Retrieved from https://github.com/CSSEGISandData/COVID-19

[13] Yousefinaghani, S., Dara, R., Mubareka, S., & Sharif, S. (2021). Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada. Frontiers in Public Health, 9. https://doi.org/10.3389/fpubh.2021.656635