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

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

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.

LITERATURE SURVEY

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.

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.

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.

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.

DATA INFORMATION

Our project will use the COVID-19 Twitter chatter dataset [10]. This is a data set that has filtered out all COVID-19 related tweets and stored their tweet id in terms of day. We will then use twitter's official Api to get detailed information such as content and location of these COVID-19 related tweets. Different NLP techniques will be performed on the content of these tweets to extract features that could be later used to predict the development of COVID-19. We intend to use four months of tweet data from 2020-08-01 to 2020-12-01, among which the first 3 months are used to train the last month is used to test and evaluate. (Considering the volume of the dataset, we will randomly sample 50,000 tweets each day. Also, we decide to only consider the situation in the United States, so tweets in other locations will be filtered out.)

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

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.

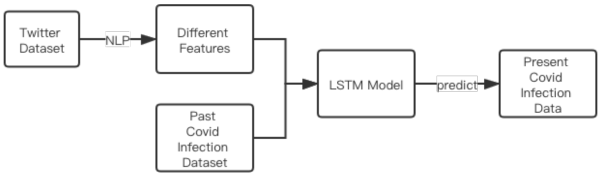

Figure 1 Approach Process Diagram

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.

EXPECTED ACCOMPLISHMENT

Our project is expected to deliver a systematic approach that leverages twitter data to enhance COVID-19 confirmed cases prediction. We should be able to first process large amounts of tweets and apply Natural Language Processing techniques to generate usable features and then fine-tune the deep learning model LSTM to obtain accurate predictions. This approach will allow real-time public reaction scattered across the internet to become useful for understanding the development of a pandemic, and it should be generalizable to other diseases and problems.

TIMELINE & RESPONSIBILITY

Stage	Task	Time	Involved members
0.0	Dataset search, idea generation, literature survey	Sep 25 – Oct 1	Everyone
0.1	Project proposal	Oct 2 - Oct 5	Everyone
1.0	Data gathering/preprocessing-Twitter	Oct 6 – Oct 15	Siyan Cai Shichen Li
1.0	Data gathering/preprocessing-COVID	Oct 6 – Oct 15	Tongzhou Yu
1.1	Twitter data NLP – sentiment score	Oct 16 – Oct 21	Siyan Cai
1.1	Twitter data NLP – emotions	Oct 16 – Oct 21	Shichen Li
1.1	Twitter data NLP – topic/term count	Oct 16 – Oct 21	Tongzhou Yu
1.2	Data concatenation	Oct 22	Everyone
1.3	Build and train LSTM model – baseline + 1 feature	Oct 23 – Oct 29	Everyone
2.0	Milestone Report	Oct 30 - Nov 2	Everyone
2.1	Train other models with more features	Nov 2 – Nov 16	Everyone
2.2	Model evaluation	Nov 16 - Nov 23	Everyone
3.0	Final Report	Nov 23	Everyone
3.1	Presentation	Nov 23/25	Everyone

Table 1 Timeline & Responsibility

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] Moore. (2020, October 14). COVID-19 Complete Twitter Dataset (daily updates). Retrieved from https://www.kaggle.com/imoore/COVID19-complete-twitter-dataset-daily-updates

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