Forecasting the COVID-19 trend based on the combination of social media and time series information

Hanyu Lu
hlu332@gatech.edu
Georgia Institute of Technology

Qiuyuan Xu
qxu317@gatech.edu
Georgia Institute of Technology

ABSTRACT

COVID-19, also known as Coronavirus, has spread throughout the world, leading to millions of confirmed cases in more than 200 countries. As such, different studies have been released to predict the growth of the disease by using different methods, which provide powerful assistance for controlling the spread of the virus. In this process, forecasting is based on diverse techniques and various parameters. However, few studies were explored based on geolocation and using text data as a variable. Our project will mainly focus on solving previous problems. First, we would like to explore whether people in different areas take different attitudes towards COVID-19 and how their attitudes changes based on sentiment analysis of Twitter data. What's more, we are eager to find out whether people's attitudes effect the trend of COVID-19. During this process, we also plan to evaluate different forecasting techniques and discuss if Tweets data is associated with forecast accuracy. In this work, we will try several popular forecasting approaches and find the one which best fits the growth of the disease. For input, in addition, to use general statistical data, we want to take text data accessed from social media(Twitter) into consideration. By using sentiment analysis, we will convert text information to mathematical vectors as input in order to explore its impact on forecast accuracy.

KEYWORDS

COVID-19, NLP, Forecast, Sentiment analysis, Machine learning

ACM Reference Format:
Hanyu Lu and Qiuyuan Xu. 2021. Forecasting the COVID-19 trend based on the combination of social media and time series information. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The COVID-19 has been spread across the whole world and has triggered a significant public health crisis. What's more, it also caused other problems, like economic crisis and decrease of the international trade. People are not only worried about getting the disease but also get involved in some mental health issues, especially when they get more information from social media. During this process, social media involvement and interaction increase dynamically. More people share their viewpoints and aspects about COVID-19 on the social media. It is easy for them to feel pessimistic when viewing the negative news on Twitter. By analyzing user-generated content on social media, such as Twitter, we can know more about the public's attitudes, thoughts, and sentiments on health status, concerns, panic, and awareness related to COVID-19. It is possible that the public's attitudes assist in predicting the the trend of COVID-19. Moreover, the attitudes can ultimately play an important role in developing health intervention strategies and design effective campaigns based on public perceptions. Based on this, we can do sentiment analysis to predict the COVID-19 trend. What's more, other information, like Google Trends, Twitter Social Mobility Index may also play a vital role in forecasting the COVID-19 trend. So our goal is to forecast the COVID-19 trend based on the above information.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 REPLY TO THE COMMENTS

1. We read some paper from top journal, including Nature Human Behavior. And we summarize them in Literature Survey Part.
2. We have already looked at the package by Noah Smith (UW)'s group on analyzing Twitter data. It is a fast and robust Java-based tokenizer and part-of-speech tagger for tweets. When we input a tweet, TweeboParser predicts its syntactic structure, represented by unlabeled dependencies. And based on the work what we have done, we plan how to use the package in our future work for another sentiment analysis method.
3. We also assume a hypothesis: positive sentiment is correlated with lower epidemic incidence.

3 LITERATURE SURVEY

Related to COVID-19, many scholars have done research on various machine learning models to forecast the spread of the global pandemic.

Iman Rahimi et al.[6] present a brief analysis of methods used in studies against COVID-19. The most popular approaches researchers have addressed to predict the spike of COVID-19 are epidemic models. Deep learning models also have the most contributions. In forecasting, confirmed cases are one of the top criteria scholars prefer to use, and some hybrid algorithms can enhance the power of forecasting methods.

Gitanjali R. Shinde et al. [8] also review and category the most popular forecasting techniques for predicting the spread of COVID-19. Different from the previous scholar, besides discussing the challenges when using the prediction approach, they summarized commonly used parameters and classified both the forecasting methods and parameters. This classification helps give researchers a more concise overview of the methods for predicting the trend of the virus.

Pooja Mehta et al.[10] proposed an algorithm for stock price prediction which takes public sentiment into account apart from mathematical parameters such as stock price. LSTM obtained the maximum accuracy which was about 92%, shown that combining sentiment with the historical stock price could reach to a stock price prediction accuracy tool.

Binrong Wu et al.[13], instead, focused on the impact of social media data on oil price forecasting, indicated that online media information such as online oil news contributes to the oil price, production and consumption prediction. Results showed that the forecasting performance with text features, financial features, and historical data is significantly improved than the forecasting performance of using only historical oil price.

Yogev Matalon[9] underscores that Opinion Inversion phenomenon plays an important role in political communication on social media, which could be used to optimize content propagation. For feature selection, they considered four prediction models: Logistic Regression, Artificial Neural Network, Random Forest, and XGBoost.

Similarly, data derived from social media against COVID-19 can also help investigate society's attitude about the current pandemic and uncover the hidden dynamics of an emerging outbreak, which gives birth to digital disease surveillance.

Kathy Lee et al.[7] designed a model which combines social media data and historical data sources to achieve more accurate forecasting at least a week ahead of time. This research used tweets that mentioned 'flu' and applied text analysis techniques to disambiguation for distinguishing if the tweet was about his/her own or someone else.

Cuihua Shen et al.[12] collected Weibo data related to COVID-19 to forecast confirmed case counts in mainland China. According to the performance of different machine learning models in classifying sick posts, the random forest algorithm earned the highest F1 score. The result showed that sick posts from social media could forecast new cases 2 to 8 days in advance outside Hubei, and up to 19 days in advance in Hubei.

I Kit Cheng et al.[1] developed a prototype early-warning system using the distribution of total tweet volume to investigate into localized outbreak predictions. Samira et al.[2] created a twitter-based data analysis framework to automatically monitor avian influenza outbreaks in a real-time manner.

Faheem Aslam et al.[3] extracted and classified the news related to COVID-19 and the result of his study can be weaved into important implications for emotional wellbeing and economic perspective.

Matthew D.R et al.[11] found that positivity is only one facet of individuals' opinions and emotionality from the very same reviews on the social media such as Amazon and Yelp offers a consistent diagnostic signal.

Krishna C.Bathina et al.[4] discovered that the language of individuals with a diagnosis of depression on social media prefer to have higher levels of distorted thinking. And online language patterns are indicative of depression-related distored thinking.

4 WHAT WE HAVE DONE

4.1 Description of dataset

We plan to use both social media data and time-series data in our project and combine them into a forecasting model.

The first data source we are going to use is the Twitter dataset, which contains tens of millions of tweets related to COVID-19 collected starting on February 6, 2020. The dataset is collected by a team of academic researchers who utilize web and social media data to study public health. And the tweets which contain keywords related to COVID-19 are all collected by Twitter public keyword streaming API. The keywords(case-insensitive) included in this collection are coronavirus, Wuhan, 2019ncov, sars, mers, 2019-ncov, wuflu, COVID-19, COVID19, COVID, covid-19, covid19, covid, SARS2, SARSCOV19 and also hashtags, like #covid19. What's more, the data also includes the information of location, which is inferred by Carmen, a geolocation toolkit. We can get the whole tweet data by tweet_id, which includes the text content, created time, location, and so on.

The diagram below shows that the number of tweets written in English in a week during the first stage of the epidemic outbreak. In February 2020, the global epidemic has just began to outbreak, and most of the cases were discovered in China. Judging from the Twitter data, although it temporarily aroused discussion, the popularity quickly dropped since there were few confirmed cases outside of China, it did not draw more attention of people.

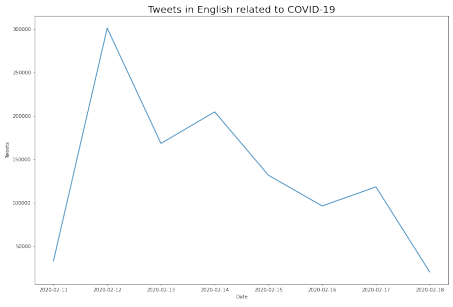

Figure 1: Tweets in English related to COVID-19

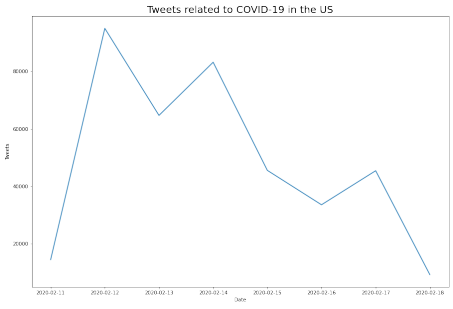

Figure 2: Tweets in US related to COVID-19

Another data source we will use is time-series data, which is Google trend. Google trends can help check the popularity and

search trends of a certain keyword within a certain time frame, mainly come from Google Search, Google Shopping, YouTube, Google News, and Google Images.

The diagram below shows the popularity and search trends of keyword "COIVD-19". It can be seen that at the peak of the global epidemic, the discussion of the epidemic was pretty high.

Figure 3: Google trends related to COVID-19

But as time goes by, the popularity of the discussion has gradually declined and tended to be flat.

The diagram below shows the words relatd to "COVID-19" which appear the most. By doing this, we could see what content people interest the most.

Figure 4: Word frequency related to COVID-19

4.2 Proposed Methodology

4.2.1 sentiment analysis method 1.

For the first method of sentiment analysis, we use textblob which is a Python library for processing textual data. The sentiment property returns the score of polarity. The score is a float within the range of -1.0 and 1.0. -1 indicates negative sentiment and +1 indicates positive sentiments.

The diagram below shows the result of sentiment analysis from 2020-02-11 to 2020-02-19. As it is shown, the majority sentiment at that time is neutral.

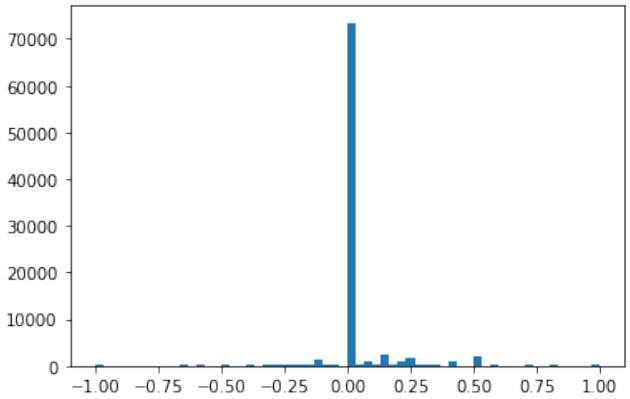

Figure 5: Sentiment analysis result

The diagram below shows the sentiment from 2020-02-11 to 2020-02-19 related to COVID-19 except the neutral sentiment. As it is shown, the emotions are relatively stable. In general, most people held a positive attitude towards Covid-19, probably because the epidemic had not spread widely around the world at that time.

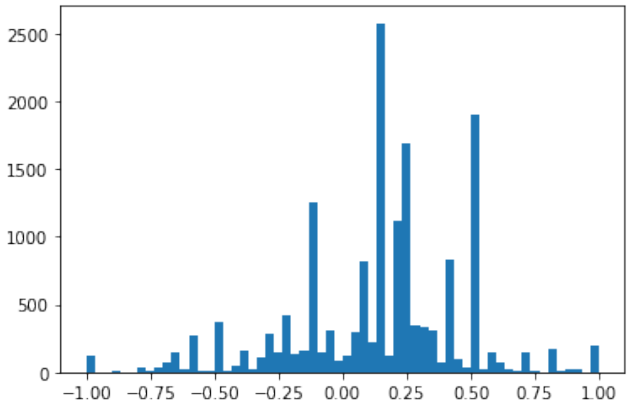

Figure 6: Sentiment analysis result except neutral

4.2.2 sentiment analysis method 2.

For the second method of sentiment analysis, there are several steps we should do.

First, we should preprocess our dataset. Breaking the sentences into each individual word is helpful. This method is called tokenization, which is efficient and convenient for computers to analyze the text data by examines what words appear in an article and how many times these words appear, and is sufficient to give insightful results.

In this step, we use tools from NLTK, Spacy to preprocess the tweets. After tokenization, each tweet transform into a list of words. Take a tweet from 2020-02-12 as an example. Its text is "Scientists: The coronavirus can only be spread by human". After tokenization, it split into several words and punctuation.

	text	lemma	POS	explain	stopword
0	Scientists	scientist	NOUN	noun	False
1	:	:	PUNCT	punctuation	False
2	The	the	DET	determiner	True
3	coronavirus	coronavirus	NOUN	noun	False
4	can	can	AUX	auxiliary	True
5	only	only	ADV	adverb	True
6	be	be	AUX	auxiliary	True
7	spread	spread	VERB	verb	False
8	by	by	ADP	adposition	True
9	human	human	NOUN	noun	False

Figure 7: Tokenization

The next step is to remove useless information and eliminate URLs. Also, symbols, digits, punctuations, @username, hashtags may also be useless information. After that, the tweet is much cleaner, but we still see some words we do not desire, for example, "and", "we", etc. The next step is to remove the meaningless words called stopwords and also. It is also necessary to remove grammar tense and transform each word into its original form when we need to use the term frequency.

In this step, we use spacy combined with regex to remove the worthless characters, and use NLTK library to remove stopwords. Take a tweet from 2020-02-12 as an example. Its content is "2015 new confirmed cases of novel coronavirus infection and 97 new deaths from 31 provincial-level regions Coronavirus CoronavirusOutbreak". The final cleaned data is ['new', 'confirm', 'case', 'of', 'novel', 'coronavirus', 'infection', 'and', 'new', 'death', 'from', 'provincial', 'level', 'region', 'Coronavirus', 'CoronavirusOutbreak'].

Then we need to transfer the tweet text into the vector. There are several ways to solve the problem, such as one-hot encoding and bag of words(BOW), TF-IDF, word2vec, etc. Among these methods, TF-IDF and word2vec are popular. TF-IDF takes into account the word counts in each document and the occurrence of this word in other documents. And Word2vec is a two-layer neural net that processes text by vectorizing words. Its input is a text corpus and its output is a set of vectors: feature vectors that represent words in that corpus.

In this step, we use word2vec to transfer tweet text into the vector.

```
-1.13947198e-02  9.70463450e-03  1.01055284e-02 -7.18149138e-04
 6.09264548e-03  1.26482799e-02 -2.96350045e-03  4.59850409e-04
-2.45545074e-03  1.06535865e-03 -1.68944441e-02 -3.25361458e-03
-5.90224772e-04 -3.99206469e-03  4.58528627e-03 -9.47209544e-03
 3.59434586e-03 -1.71881151e-02  2.26734208e-03 -4.69123634e-03
 1.41825993e-02 -1.93609348e-02 -5.04984454e-03  9.27978828e-04
-6.07692788e-03  2.01499179e-04  1.55134240e-02 -1.99119111e-03
 3.07294546e-03 -4.55237211e-03  3.18397850e-03 -1.50990541e-03
 2.27438335e-03  4.16283750e-03  1.02250543e-02 -7.81685047e-03
 1.06133036e-02  7.53787383e-03  1.32766387e-02  1.99425310e-02
 1.47109292e-03  1.59035146e-03  1.23895913e-02  2.04978570e-03
 1.16985554e-02  4.87029116e-04 -1.51719658e-02  3.58823001e-03
 1.10500081e-02  1.35722452e-02 -1.12156325e-02 -2.66857176e-04
```

Figure 8: Part of tweet text vector

5 WHAT WE WILL DO

5.1 Data Analysis

Now, because the number of everyday Twitter data about COVID-19 is more than 100 thousand and getting the data from Twitter takes a lot of time, we only show a short period of Twitter data analysis in this stage. In the next period of time, our codes will automatically extract Twitter data every day and store data. We will carry out data cleaning, data statistical analysis.

5.2 Sentiment Analysis and Prediction

Now, we use the package of textblob which can directly convert Twitter text data into sentiment ploarity values. But we also try another way to do sentiment analysis. The main idea is to convert text data to vector and so sentiment analysis based on machine learning models.

For second way to do sentiment analysis, we have already fininshed the steps of tokenization and removed useless information and converted to data into vector with the help of Word2vec.

In the next period, we may try TF-IDF to convert text data into vector and use different models to do sentiment analysis such as SVM, KNN and Random Forest. SVM is a supervised learning method used for classification, regression, and outliers detection. For KNN, we calculate the distance and determine the classification in which data will be in. Random forests operates by constructing a set of decision trees at training time. After we do the sentiment analysis, we can use the result to predict the COVID-19 trend. we will try different models, such as BPNN, RNN and LSTM.

BPNN is a method whose network of neurons can be trained with a training dataset in which output is compared with desired output and error is propagated back to the input.RNN is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. LSTM is a special recurrent neural network that has feedback connections. Different models may suitable for the different datasets. So we are eager to try several models to get more information and compare their performance in predicting the COVID-19 trend.

5.3 Evaluation

For the sentiment analysis part, we can calculate the Accuracy, Precision, Recall, and F-value.

Accuracy is an evaluation metric that allows you to measure the total number of predictions a model gets right. Precision evaluates how precise a model is in predicting positive labels. Recall calculates the percentage of actual positives a model correctly identified (True Positive), when the cost of a false negative is high, we should use recall. F_1 Score is the weighted average of Precision and Recall which takes both false positives and false negatives into account. Those metrics can be calculated based on the table shown below.

Table 1: Classification

	Class 1	Class 2
Identified as Class 1	a	b
Identified as Class 2	c	d

$$Accuracy = \frac{a+d}{a+b+c+d}$$
$$Precision = \frac{a}{a+b}$$
$$Recall = \frac{a}{a+c}$$
$$F_1score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

For predicting part, we can evaluate by calculating the MAE, MSE and RMSE[5].

MAE, known as the mean absolute error, represents the average of the absolute difference between the actual and predicted values in the dataset. MSE, known as mean squared error, represents the average of the squared difference between the original and predicted values in the data set. Root Mean Squared Error is the square root of Mean Squared error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}|$$
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y})^2$$
$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y})^2}$$

To measure the success, the less the MSE, the larger the F_1 score, the better we predict.

5.4 Expected Result

In this project, we want to get conclusions from three aspects.

1. With the help of NLP techniques, after investigating social media text data by using sentiment analysis, we want to test our hypothesis. According to previous surveys, social media information is a powerfully related toCOVID-19 trend. Therefore, we expect that social media data contributes to the accuracy of forecasting.

2. In the survey, we learned that there exists a variety of forecasting machine learning and deep learning models to predict the future trend of coronavirus. Based on the data we have, including Twitter data and official time-series data, we expect to compare the most accurate forecasting machine learning and deep learning models among RNN, LSTM and BPNN, and find the one which best fit the data perfectly and choose it to complete the prediction of the future spike of the virus. To measure the success, the less the MSE

3. In our work, we will make full use of the geolocation information in the datasets to detect people's attitude changes in different location. We expect to explore if different regions will be suitable for different forecasting models. Relying on that, we may make targeted predictions on the growth of the epidemic in different regions.

As we see in the survey, different studies have been released to successfully predict the growth of the disease by using different methods. However, few studies were explored based on geolocation and using text data as a parameter. By doing this project, we may improve the accuracy of prediction by considering the difference of geolocation and social media information. Then, based on the conclusion we obtain from this project, we can assist in developing health intervention strategies and design effective campaigns based on public perceptions.

REFERENCES

[1] Quercia D. Zhou K. et al Aiello, L.M. 2021. How epidemic psychology works on Twitter: evolution of responses to the COVID-19 pandemic in the U.S. *Humanit Soc Sci Commun 8* (2021). https://doi.org/10.1057/s41599-021-00861-3

[2] Quercia D. Zhou K. et al. Aiello, L.M. 2021. How epidemic psychology works on Twitter: evolution of responses to the COVID-19 pandemic in the U.S. *Humanit Soc Sci Commun 8* (2021). https://doi.org/10.1057/s41599-021-00861-3

[3] Awan T.M. Syed J.H. et al. Aslam, F. 2020. Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak. *Humanit Soc Sci Commun 7* (2020). https://doi.org/10.1057/s41599-020-0523-3

[4] ten Thij M. Lorenzo-Luaces L. et al Bathina, K.C. 2021. Individuals with depression express more distorted thinking on social media. *Nat Hum Behav 5* (2021). https://doi.org/10.1038/s41562-021-01050-7

[5] Alvin Wei Ze Chew, Yue Pan, Ying Wang, and Limao Zhang. 2021. Hybrid deep learning of social media big data for predicting the evolution of COVID-19 transmission. *Knowledge-Based Systems* 233 (Dec. 2021). https://doi.org/10.1016/j.knosys.2021.107417

[6] Rahimi I., Chen F., and Gandomi. 2021. A review on COVID-19 forecasting models. *Neural Comput Applic* (April 2021). https://doi.org/10.1007/s00521-020-05626-8

[7] Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2017. Forecasting Influenza Levels Using Real-Time Social Media Streams. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. 409–414. https://doi.org/10.1109/ICHI.2017.68

[8] Parikshit Mahalle, Asmita B. Kalamkar, Nilanjan Dey, Jyotismita Chaki, Aboul ella Hassanien, and Gitanjali R. Shinde. 2020. Forecasting Models for Coronavirus (COVID-19): A Survey of the State-of-the-Art. *TechRxiv* (April 2020). https://doi.org/10.36227/techrxiv.12101547.v1

[9] Magdaci O. Almozlino A. et al. Matalon, Y. 2021. Using sentiment analysis to predict opinion inversion in Tweets of political communication. *Sci Rep 11* (2021). https://doi.org/10.1038/s41598-021-86510-w

[10] Mehta P, Pandya S, and Kotecha K. 2021. Harvesting social media sentiment analysis to enhance stock market prediction using deep learning. *PeerJ Computer Science* (April 2021). https://doi.org/7:e476

[11] Rucker D.D. Nordgren L.F Rocklage, M.D. 2021. Mass-scale emotionality reveals human behaviour and marketplace success. *Nat Hum Behav 5* (2021). https://doi.org/10.1038/s41562-021-01098-5

[12] Luo C Zhang J Feng B Liao W Shen C, Chen A. 2020. Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study. *J Med Internet Res* (May 2020). https://doi.org/10.2196/19421

[13] Binrong Wu, Lin Wang, Sirui Wang, and Yu-Rong Zeng. 2021. Forecasting the U.S. oil markets based on social media information during the COVID-19 pandemic. *Energy* 226 (July 2021). https://doi.org/10.1016/j.energy.2021.120403