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. Forecasting is based on diverse techniques and various parameters. However, few studies were explored based on geolocation and using text data as a parameter. Our project will mainly focus on solving previous problems. In this project, based on geolocation we 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 as parameters, we want to take text data accessed from social media(Twitter) into consideration. By using sentiment analysis, we will convert text information to mathematical parameters 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, 4 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. During the process, 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. 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, which can ultimately assist 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 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.[2] 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 especially SIR, SEIR models. Deep learning models also have the most contributions. But those methods show pros and cons, for instance, even though results obtained through deep learning can be comparable to human expert performance, the training process is expensive. 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. [4] 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. Forecasting techniques can be categorized into mathematical models and machine learning techniques. Mathematical models which are efficient and traditional approaches were widely used in past pandemics. In recent years, data science methods are more and more popular due to their accuracy, especially machine learning techniques. However, there are obvious challenges from choosing appropriate parameters, selecting a suitable ML model to train the model. For analysis, diverse kinds of parameters are taken into consideration. The categorization for parameters is based on the databases it comes from. Data accessed from the authorized organization, such as WHO, is mainly used as a statistical and mathematical parameter. Major significance is the number of daily deaths, mobility, transmission rate. Data obtained from communication media includes not only the quantitative indicators but also non-quantitative text information. For example, the internet searches government policies and related news. This classification helps give researchers a more concise overview of the methods for predicting the trend of the virus.

These two reviews give us a hint of how to choose the proper machine learning model and considering the impact of parameters on the performance of forecasting.

2.1 Prediction Using Social Media Data

Predictions based on the information accessed from social media have been applied in many areas.

Pooja Mehta et al.[5] proposed an algorithm for stock price prediction which takes public sentiment into account apart from mathematical parameters such as stock price. The analysis used data sets collected from various social media sources such as IIFL, Economic Times and from historical numerical stock data intended to predict the Indian SM for a period of time. To quantify user's sentiment, defined 0 as neutral sentiment, < 0 as negative sentiment, > 0 as positive sentiment. After data pre-processing, different machine learning methods were used to combine sentiment expectation with SM numerical data. In the process of prediction, among several classification approaches such as Linear regression, Naive Bayes technique, Maximum entropy, Decision tree and LSTM, 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.[7], 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. This study applied the neural network(CNN) model to extract information in oil news. Then input text features, financial market data, and historical oil data into several forecasting techniques, including backpropagation neural networks (BPNN), support vector machines (SVM), multiple linear regression (MLR), recurrent neural network (RNN), and long short-term memory (LSTM). Results showed that the BPNN model obtains the best forecasting performance, what's more, the forecasting performance with text features, financial features, and historical data is significantly improved than the forecasting performance of using only historical oil price.

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. Google Flu Trends is one of the best successful examples of digit disease surveillance applications. Generally, data generated from social media activities can be a powerful supplement for disease prevalence prediction.

Kathy Lee et al.[3] 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. By experimenting with different combinations of CDC and Twitter data, the best prediction for the current flu level was the latest 3 weeks' CDC plus the latest 5 week's Twitter data with a correlation coefficient of 0.9525, performance improved compared with the latest 3 weeks' CDC data. The same improvement happened for the best 1-week ahead prediction mode. It clearly shows that compared to using only historical CDC data, combining twitter data with CDC data improves the forecasting performance. For predictive modeling algorithms, multilayer perceptrons (MLP) had the best performance, so 3-layer MLP with 4 activation units in the hidden layer was used in this study. As the result, the proposed model could not only estimate the current flu level 2 weeks ahead of CDC data but also predict future flu activity 3 weeks ahead with high accuracy.

Cuihua Shen et al.[6] collected Weibo data related to COVID-19 to forecast confirmed case counts in mainland China. After obtaining COVID-19 related posts, the author used supervised machine learning algorithms to identify sick posts. According to the performance of different machine learning models in classifying sick posts, the random forest algorithm earned the highest F1 score. To estimate the final models, ordinary least squares regression with robust standard errors was used and 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.

3 PLAN OF ACTIONS

3.1 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. It provides three levels of location information: country, state, and city. If the tweet has a place or coordinates field, Carmen returns this information. In conclusion, the team provides the following information to help us get the data. The information is as follows:

Variable	Format	Description
tweet id	int	The ID of the tweet to download from Twitter
user_id	int	The user ID of the author of this tweet
date	string	Standard Twitter date format
keywords	list	COVID-19 related keywords used to identify the tweet
location	dictionary	Include country, state, and city

Figure 1: Tweet Record

So, based on that, we can get the whole tweet data by tweet_id, which includes the text content, created time, retweet times, location, and so on.

Another data source we will use is time-series data, which include the NYT dataset, Google trend, and Twitter Social Mobility Index.

NYT dataset is for the daily and historical cumulative cases and deaths count in each county and state in the U.S. since the beginning of the pandemic. For historical data, it contains the number of daily cases and deaths. The counts include both laboratory-confirmed

cases and probable cases, which are based on a combination of evidence from clinical, epidemiologic, or serological testing or from vital records, but without a confirmatory laboratory RNA test. For live data, in addition to the information contained in the historical data, it also includes detail on the number of confirmed and probable cases separately. And both historical and live data are released in three levels: U.S., states, and counties. The main difference between the historical and live data is that the numbers in the historical files are the final counts at the end of each day, while the live data may be a partial count released during the day but may not be the final.

Live data

Variable	Format	Description
date	string	Standard date format
cases	int	The number of cases from COVID-19, including bothe confirmed and probable
deaths	int	The number of deaths from COVID-19, including both confirmed and probable
confirmed_cases	int	The number of laboratory confirmed Covid-19 cases only, or blank if not available.
confirmed_deaths	int	The number of laboratory confirmed Covid-19 deaths only, or blank if not available.
probable_cases	int	The number of probable Covid-19 cases only, or blank if not available.
probable_deaths	int	The number of probable Covid-19 deaths only, or blank if not available.
county	string	Name of the county
state	string	Name of the state
fips	int	A standard geographic identifier

Example

National-Level Data

	date	cases	deaths	confirmed_cases	confirmed_deaths	probable_cases	probable_deaths
1	2021-10-04	43654161	701326				

County-Level Data

	date	county	state	fips	cases	deaths	confirmed_cases	confirmed_deaths	probable_cases	probable_deaths
1	2021-10-04	Autauga	Alabama	01001	9780	142	8186	126	1594	16

State-Level Data

	date	county	state	fips	cases	deaths	confirmed_cases	confirmed_deaths	probable_cases	probable_deaths
1	2021-10-04	Autauga	Alabama	01001	9780	142	8186	126	1594	16

Figure 2: NYT-Live Data

Historical data		
Variable	Format	Description
date	string	Standard date format
cases	int	The number of cases from COVID-19, including bothe confirmed and probable
deaths	int	The number of deaths from COVID-19, including both confirmed and probable
county	string	Name of the county
state	string	Name of the state
fips	int	A standard geographic identifier

Example		
National-level Data	State-Level Data	County-Level Data
date,cases,deaths 2020-01-21,1,0 ...	date,state,fips,cases,deaths 2020-01-21,Washington,53,1,0 ...	date,county,state,fips,cases,deaths 2020-01-21,Snohomish,Washington,53061,1,0 ...

Figure 3: NYT-Historical Data

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.

Twitter social mobility index is a measure of social distancing and travel derived from Twitter data, which could help measure how much a user travels in a given week.

- Location: The US state, city, or territory and the US as a whole.
- mobility_before_distancing: The social mobility index for the entire time period of the dataset before social distancing.
- mobility_after_distancing: The social mobility index after social distancing.
- reduction: The percent reduction of social mobility.
- num_users: The number of unique Twitter users in this location that were used to compute the metric.
- num_records: The number of unique Tweets in this location that were used to compute the metric.

3.2 Proposed Methodology

3.2.1 sentiment analysis.

In the section on the sentimental 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.

The next step is to remove useless information and eliminate URLs. Also, symbols, digits, punctuations, @username, hashtags may also be useless information.

After the above steps, the news article 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.

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.

We may choose different models to do sentiment analysis such as support vector machines(SVM), K nearest neighbors(KNN), and Random Forest. SVM is a supervised learning method used for classification, regression, and outliers detection. The goal of the SVM is to find a hyperplane that can distinctly classify the data points. KNN can be used for classification. It assumes similar things exist in close proximity. By calculating the distance, we can determine the classification in which data will be in. Random forests can be used for classification, regression, and other tasks. It operates by constructing a set of decision trees at training time.

3.2.2 Prediction of COVID-19.

After we do the sentiment analysis, we can use the result to predict the COVID-19 trend. we will try different models, including SVM and neural networks, such as Back-Propagation Neural Network(BPNN), Recurrent Neural Network(RNN) and Long short-term memory(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. A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. Long short-term memory (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.

3.3 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

$$Accuracy = \frac{a+d}{a+b+c+d}$$

$$Precision = \frac{a}{a+b}$$

$$Recall = \frac{a}{a+c}$$

$$F_1 score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

For predicting part, we can evaluate by calculating the MAE, MSE and RMSE[1].

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.

3.4 Expected Result

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

1. With the help of NLP techniques, such as KNN, SVM and Random forest, after investigating social media text data by using sentiment analysis, we want to identify if adding social media data to conventional numerical data could improve the accuracy of the predictions. According to previous surveys, social media information is a powerful supplement for forecasting. 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. Different models have their own advantages and limitations. 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.

3. In our work, we will make full use of the geolocation information in the datasets to make predictions based on location differences. We expect to explore if different regions will be suitable for different forecasting models. Relying on that, we will 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.

3.5 Timeline

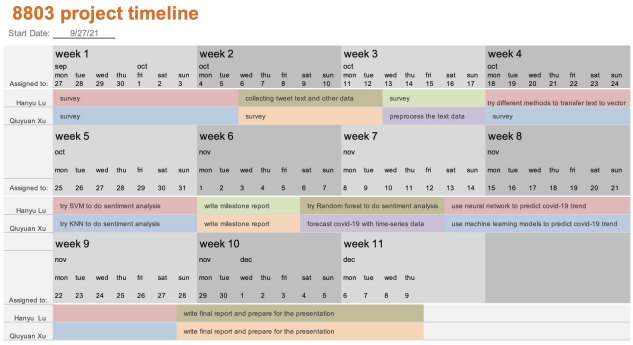

Figure 4: Timeline

REFERENCES

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] 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