Forecasting the COVID-19 trend based on the combination of social media 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 change based on sentiment analysis of Twitter data. What's more, we are eager to find out whether people's attitudes affect 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 information. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 7 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 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 trend of COVID-19. Moreover, attitudes can ultimately play an important role in developing health intervention strategies and designing 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 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.

2 PROBLEM DEFINITION

In this project, we want to understand whether social media information has an impact on the future trend of the covid epidemic, that is, whether social media information can effectively predict the development of the pandemic. In order to accomplish this goal, there are some steps we need to follow. Firstly, we should select appropriate social media information and convert it into quantifiable parameters in order to be a better fit for mathematical models prediction. Secondly, we need to compare the advantages and disadvantages between algorithms, and choose the appropriate model for prediction; Thirdly, by applying the preprocessed data into appropriate models, we can make the conclusion that whether social media data can accurately predict the trend of the epidemic.

3 RELATED WORK AND SURVEY

Related to COVID-19, many scholars have done research on various machine learning models to forecast[1] the spread of the global pandemic.

Yogev Matalon[7] 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 XG-Boost.

Similarly, data derived from social media[4] 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.[6] 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

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

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. 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.

Matthew D.R et al.[8] 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.[2] 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 PROPOSED METHOD

4.1 Intuition

In the current society, the Internet has become an indispensable tool, and social media has gradually become the main window for people to communicate. Especially during the epidemic, when people have to maintain social distance or have to self-isolate, communication and information exchange between people are carried out through social media. Therefore, it is conceivable that there is a large amount of information stored on social media, and this information contains people's attitudes and views on various events, while the public's attitude towards events is likely to change the development of this event. This inspires us that we could predict the trend of events through people's attitudes. For example, during the presidential campaign, we can predict the outcome of the campaign in advance by analyzing the public's attitudes towards different candidates on social media. Similarly, we may predict the future development of the epidemic by analyzing the public's attitudes towards the epidemic. In the past, the forecast of the epidemic was usually based on observed time series data, and few people applied network information into the forecast of the epidemic. We believe that the use of social media data may also be a good way to predict the trend of covid pandemic.

4.2 Description

(1)Approach

To extract the public's attitudes, we need to apply sentiment analysis. For doing that, 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 second method for solving text information is to convert the word to vector and match it to the dictionary. 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 examining 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 transforms 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. 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 Coronavirus Outbreak". 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. After finishing all the steps, we can choose a suitable dictionary to do the emotional analysis.

For predicting part, we can evaluate by calculating the MAE, MSE and RMSE[3].

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.

(2)Algorithms and Models To predict the trend of the Covid pandemic, we choose several prediction models including time series models, machine learning, and deep learning.

Linear Regression: A linear equation combines a set of input(x) and the solution(y) which could be predicted by the input. It assigns one scale factor to each input so when we have new input, we could use this equation to predict the new solution.

SVR: For linear regression models, the objective is to minimize the sum of squared errors. But what if we are only concerned about reducing error to a certain degree and ignore the errors which fall within the acceptable range? By using SVR, we have the flexibility to define the acceptable range in the model and find the appropriate line to fit the data. It allows us to choose acceptable error margin() and acceptable error rate.

ARIMA: ARIMA(Auto Regressive Integrated Moving Average) is a class of models to forecast future values given time series based on its own past values. Any 'non-seasonal time series that exhibits patterns and is not a random white noise can be modeled with ARIMA models.

TCN: TCN(Temporal Convolutional Network)[5], consists of dilated, causal 1D convolutional layers with the same input and output lengths. It is designed based on two basic principles. Firstly, the convolutions are causal, so there is no information leakage from future to past. To achieve it, the TCN uses causal convolutions. Secondly, the architecture can take any length of the sequence and map it to an output sequence with the same length. For the second point, the TCN uses a 1D fully-convolutional network architecture, where each hidden layer is the same length as the input layer.

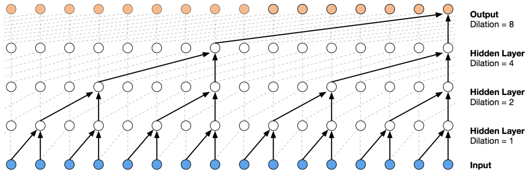

Figure 1: TCN

5 RESULTS

5.1 Which dataset we should choose

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, 2019ncov, sars, mers, 2019-ncov, 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 the result of sentiment analysis.

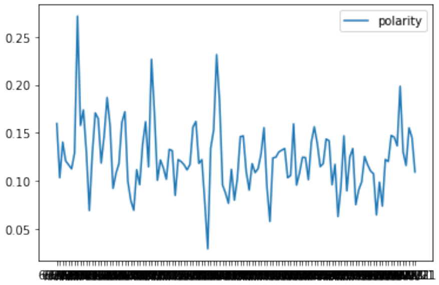

Figure 2: The trend of polarity according to the date

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 coming from Google Search, Google Shopping, YouTube, Google News, and Google Images.

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

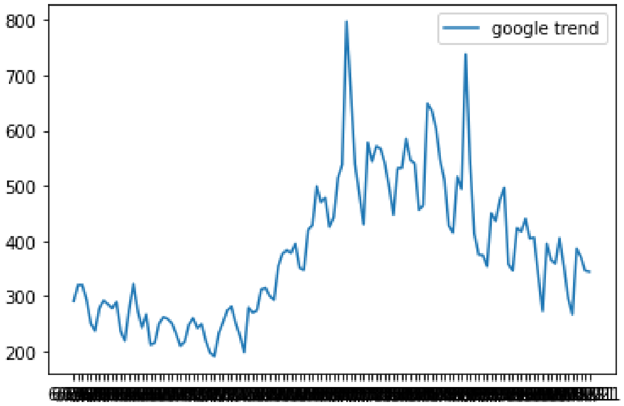

Figure 3: Google trends related to COVID-19

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

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

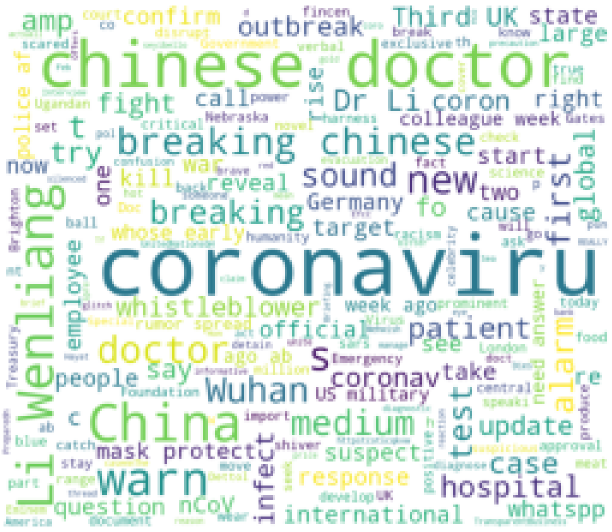

Figure 4: Word frequency related to COVID-19

For prediction, we would use the dataset from CDC to get the number of new cases.

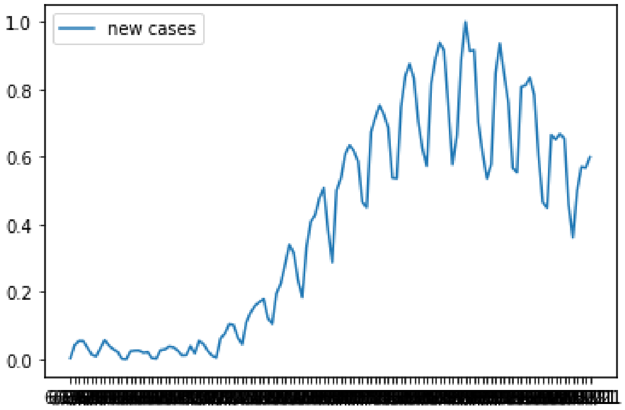

Figure 5: The number of new cases

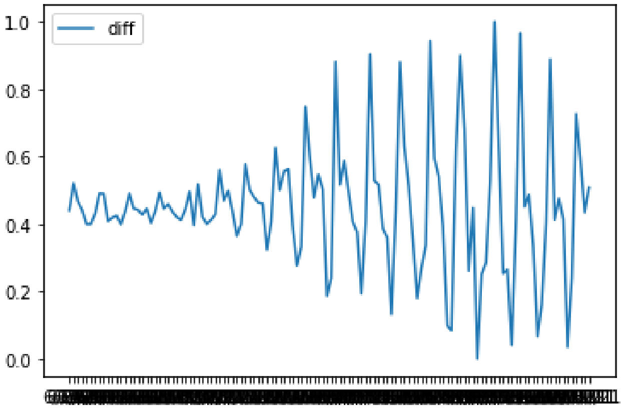

Figure 6: Trend of daily new cases

In the project, we choose the data from June 2021 to September 2021 to complete the model fitting.

5.2 Which parameter we should choose

5.2.1 Input:

Because we want to figure out the relation between social media information and the trend of Covid-19, the input values should be the data obtained from social media. According to the dataset we choose, from Twitter, we can obtain social media information by analyzing tweets. From the text of tweets, we can use sentiment analysis to get its attitude by directly using python library "textblob". Or we could convert from each word to vector and make it as the input. Comparing these two methods, we find that using "textblob" to obtain the value of polarity is more useful because it can represent the attitude of each tweet to a certain extent, changing word to vector, however, may have meaningless information.

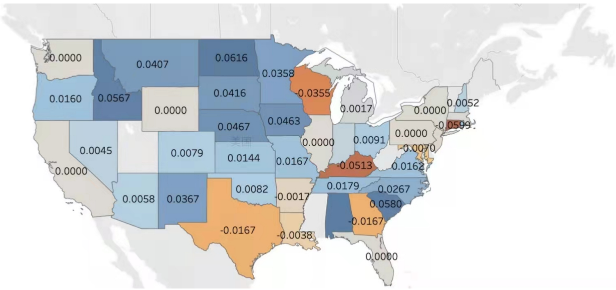

Figure 7: Sentiment analysis for each location in the US

Another dataset we choose is google trend. We select some word which is related to Covid-19 and get the frequency of each word. The value can represent the popularity of related topics.

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.

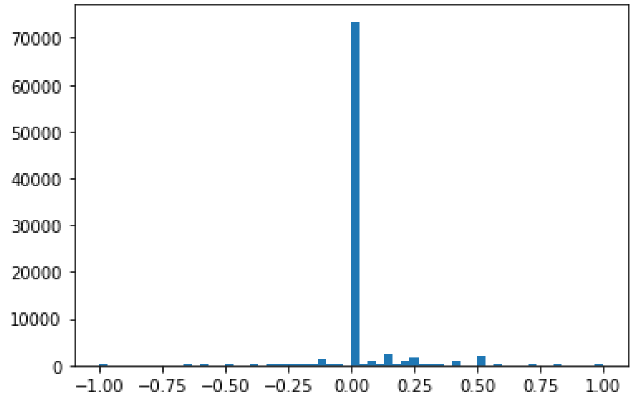

Figure 8: Sentiment analysis result

The diagram below shows the sentiment from 2020-02-11 to 2020-02-19 related to COVID-19 except for 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.

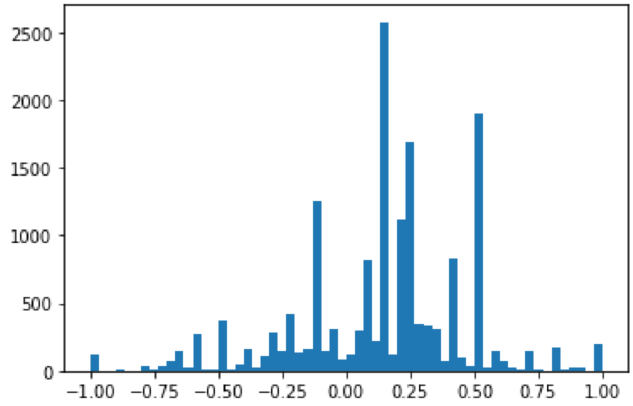

Figure 9: Sentiment analysis result except neutral

5.2.2 Output:

Generally speaking, we use the number of new cases to indicate the trend of the Covid-19. When the number of new cases increases, it indicates that the epidemic is in an up period; when the number of cases falls back, it indicates that the epidemic is leveling off. Similar to the number of new cases, we can calculate the difference in the number of new cases between the current day and the previous day, which can also be used as an indicator of the trend of the epidemic. When the number of new cases increases significantly compared to the previous day, it can indicate that the epidemic is showing an outbreak trend.

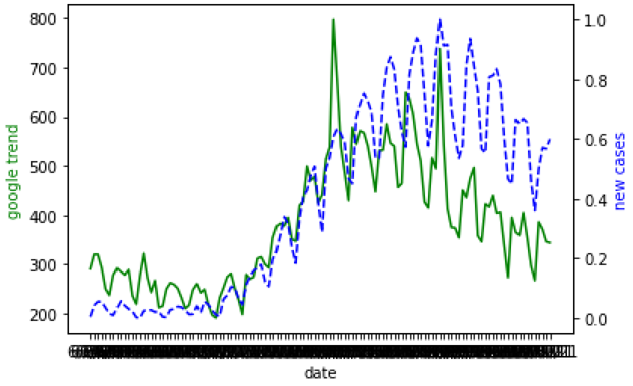

Figure 10: Comparison of the trends of google trend data and the number of new cases

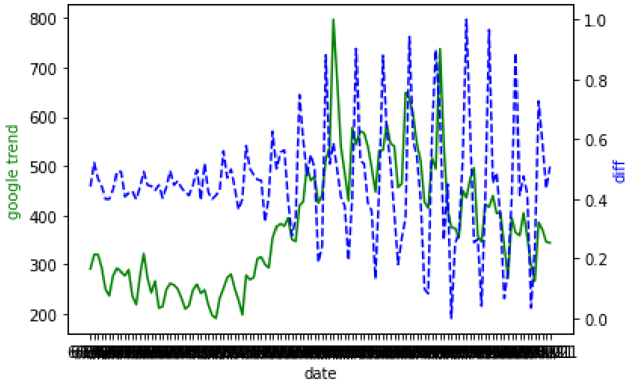

Figure 11: Comparison of the trends of google trend data and the difference in the number of new cases compared to the previous day

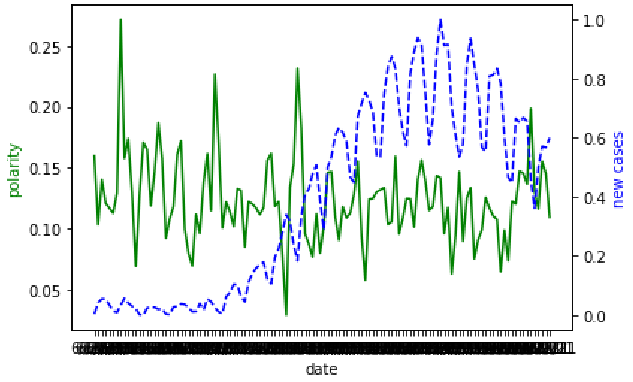

Figure 12: Comparison of the trends of polarity and the number of new cases

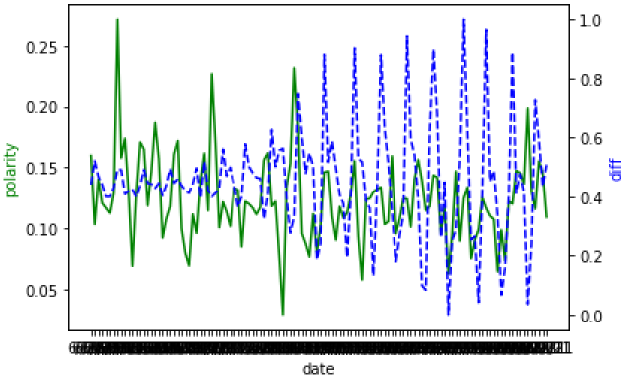

Figure 13: Comparison of the trends of polarity and the difference in the number of new cases compared to the previous day

5.3 Which model we should choose

As we talked about before, we have two different kinds of input and two different kinds of output. Our goal is to compare the prediction accuracy of different models in our dataset. And to find which model is more fit for predicting new COVID-19 cases and the difference of new cases compared with the last day. First, we choose the input, output, and model. Then, we use the first 30-day data to predict the 31st day's new cases. The 30-day data can be split into both training sets and validation sets. Then add the actual 31st data to the dataset and predict the 32nd day's new cases. By iterating this process, we can get the prediction series of the y. Meanwhile, we also calculate the validation MSE. For evaluation, we calculate, store, and each model's prediction MSE.

5.3.1 Linear Regression:

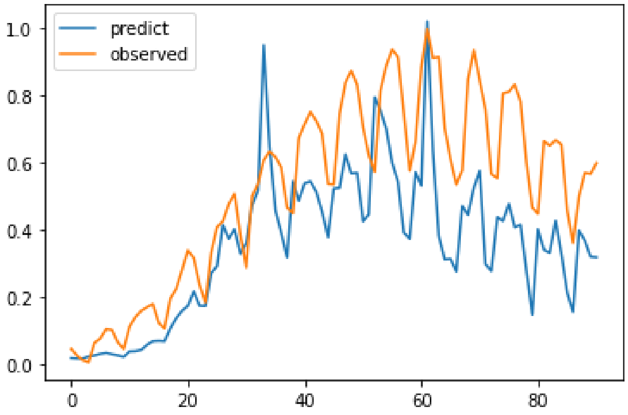

Figure 14: Result of prediction using linear regression

According to figure 8-11, google trend data and the number of new cases share a similar trend during a certain period of time, so we choose them to fit the model. After evaluation, we can calculate that MSE=0.0791660467135234, RMSE=0.281364615247766, which it's an acceptable result.

5.3.2 SVR:

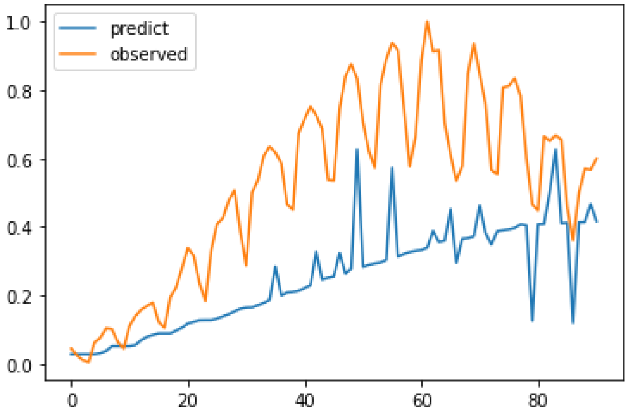

Figure 15: Result of prediction using SVR

After evaluation, MSE=0.03326636951380644 RMSE=0.18239070566727472. Even though MSE is better than the previous model, the graph shows that it does not predict well.

5.3.3 ARIMA:

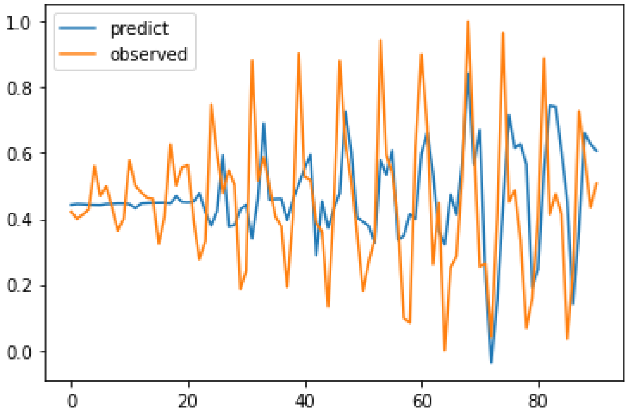

Figure 16: Result of prediction using ARIMA

The result shows that the best model for this input is ARIMA(2,0,3)(0,0,0), and MSE=0.0097457202153172, RMSE=0.09872041437978873. It seems that ARIMA model fits the input better than the previous models.

5.3.4 TCN:

First, as we talked about before, the whole trend of Google trend is similar to the trend of new cases, so we try to fit the TCN model with the dataset of both new cases and google trends.

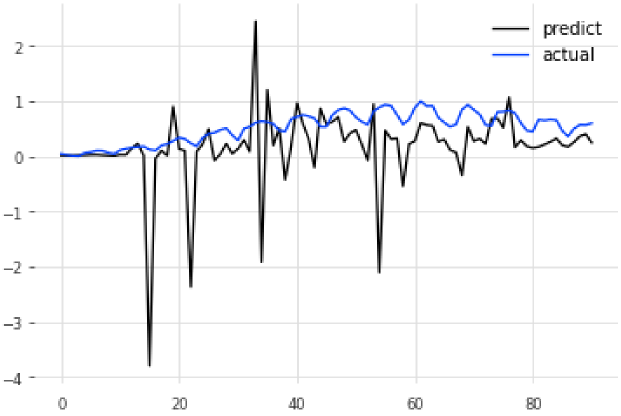

Figure 17: The prediction result of new cases using method of TCN and dataset of Google Trend

From the figure, we find the prediction result is convincing, especially in trend. The TCN model "catch" almost every "up and down". And the MSE of TCN is 0.0138

5.3.5 Evaluation

After comparing the performance of 4 different models, we found that ARIMA model has the least MSE, which means that this model predicts the value the best among the models we have selected.

6 CONCLUSION AND DISCUSSION

By comparison, it is found that different models, different input data values, and different choices all have a great influence on the final prediction results. ARIMA is good at capturing the trends of the data itself, tcn is good at capturing inflection points, and linear regression and svr are good at capturing the relationship between variables. According to our research, google trend data plays a great role in our two y predictions, which is the prediction for the trend of COVID-19. Also, it also shows that people's emotions have an impact on the predictions, but the relationship between sentiment and prediction of trend is not obvious. Instead, we found that the popularity of covid related topics has more impact on the prediction of the epidemic. It can be seen that the guidance of public opinion plays an important role in the control of the epidemic and the formulation of policies.

7 RESPONSE TO MILESTONE COMMENT

1. In the final report, we find more references which are more suitable and scientific. 2. We also clarified our problem definition, and responded to them one by one with corresponding solution.

REFERENCES

[1] Hisham Al-Mubaid and Izzat Alsmadi. 2020. Analysis and Prediction of COVID-19 Timeline and Infection Rates. In *Proceedings of the 12th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (Virtual Event, Netherlands) *(ASONAM '20)*. IEEE Press, 792–797. https://doi.org/10.1109/ASONAM49781.2020.9381338

[2] 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

[3] 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

[4] Keyan Ding, Ronggang Wang, and Shiqi Wang. 2019. Social Media Popularity Prediction: A Multiple Feature Fusion Approach with Deep Neural Networks. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2682–2686. https://doi.org/10.1145/3343031.3356062

[5] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D. Hager. 2016. Temporal Convolutional Networks: A Unified Approach to Action Segmentation. arXiv:1608.08242 [cs.CV]

[6] 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

[7] 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

[8] 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