Leveraging Internet Data and Deep Learning Method for COVID-19 Case Prediction in California

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. Besides, Google Search and Facebook are also essential internet platforms that could provide valuable data for prediction. Delphi research group at CMU [14] provides these data, we also use these data to train our model, compare their performance with tweet datasets and then try to improve performance by combing several different data sources.

2.RESPONSE TO MILESTONE COMMENT

Hypothesis testing

The original purpose of our project was to predict COVID-19 pandemic with the help of Twitter information. Beyond simply experimenting with Twitter information, we are also interested in discovering what Internet features are most helpful to COVID-19 prediction, and how we can leverage the power of them. Therefore, we didn't stop at adding Google and Facebook features but also did comparison experiments where experiments were based on Google-only data (Google features and covid data) or Twitter-only data (tweets features and covid data) or covid-only data or combinations of various internet features.

Insufficiency of features

To bring in more information into our experiments, we added Google search data of ageusia and anosmia, and Facebook survey of COVID-like symptoms into our feature pool, which surprisingly turned out to be the most powerful information we found to help predict COVID-19 pandemic.

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 California. We will focus on extracting useful insights from Internet data and finding the optimal prediction model and powerful features.

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.

COVID-19 Google and Facebook Data

Delphi group at CMU has published many covid-19 related dataset, among which we use Google and Facebook data. For Google data, we use Google search data for ageusia and anosmia. For Facebook data, we use Facebook survey results on covid-like symptom.

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

Given the cleaned twitter data, we will apply NLP techniques to extract useful features in two ways. One is sentiment analysis, which will classify tweets as positive, negative, or neutral. The proportions of these three categories might reflect public attitude towards COVID and thus function as the feature for COVID prediction. Second is frequency of keywords. The number of tweets including certain keywords might be useful when predicting COVID.

Combined with the other feature data found, 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 for comparison. Each model will be trained with and compared using the same test set. We will tune the parameters and retrain as needed. The training data will contain the first four months (August to November), while the testing data is the last month of December.

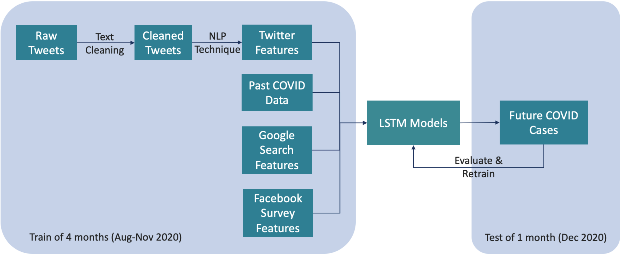

Figure 1 Approach Process Diagram

6.2 Algorithms involved

6.2.1 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.

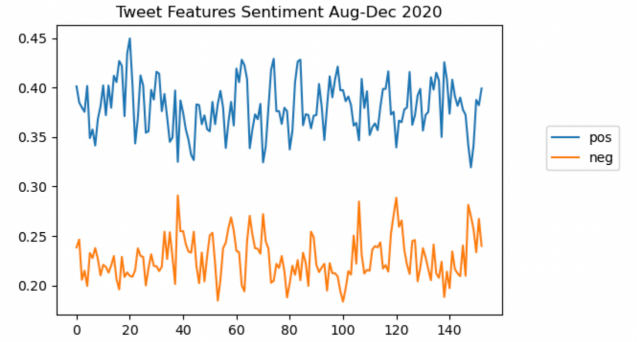

Figure 2 Tweet Features Sentiment Graph

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.

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

Table 1 Categorical Keyword List

We can observe that the frequencies of these three categorical words increases towards the last month in Figure 3.

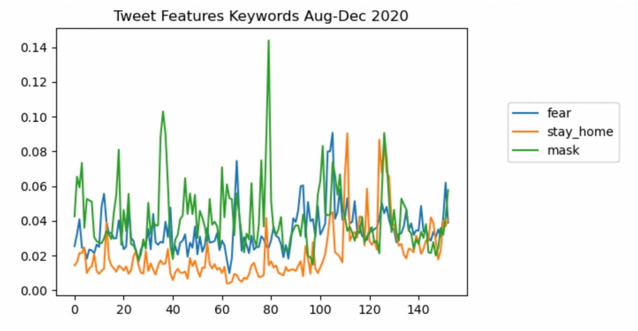

Figure 3 Tweet Features Keywords Graph

6.2.2 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 death cases and Internet information, including Google, Twitter, and Facebook features.

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.

8.EXPERIMENTS & RESULTS

8.1 Model Settings

Our project focuses on the prediction of COVID-19; thus, the most important model is the LSTM prediction model. We defined our LSTM in Keras with 50 neurons in the first hidden layer and 1 neuron in the output layer. The input shape differs according to datasets we are using which will be introduced thoroughly in next part. We used Adam as our optimizer and mean absolute error (MAE) as the loss function. For training part, the models were fitted for 100 epochs with batch size of 72.

8.2 Comparison Experiments

To show whether and how much tweets' information helped COVID prediction, we designed series of comparison experiments. However, to make the comparison valid, we kept the most model settings consistent. All models are trained by data from 08/01/2020 to 12/01/2020 and are tested on data from 12/02 to 12/31/2020.

The major difference among these experiments is the features as shown below:

1. COVID-only (baseline):
As our baseline, we only consider COVID-19 cumulative confirmed and death cases into account.

2. Twitter-only (2+5 features):
Original purpose of our project is to evaluate the effect of tweets' information in predicting COVID pandemic, thus we add 4 extra features: emotion features (positive, negative rates) and keywords features (fear, stay and mask rates), into our feature pool.

3. Google-only (2+2 features):
To tell how much improvement tweets did to the prediction, we introduced Google features (ageusia and anosmia search rates). In this case, model takes Google features and COVID features into account.

4. Facebook-only (2+1 features):
In this case, model takes Facebook features and COVID features into account.

5. Three Mixed cases:
In these cases, we combined two of above-mentioned internet resources with COVID features, i.e., Google + Facebook (2+2+1 features), Google + Twitter (2+2+5 features) and Facebook + Twitter (2+1+5 features).

6.Full-feature (2+5+2+1 features):
After all these, we combined all three groups of features with COVID features to build a more comprehensive experiment.

8.3 Results Analysis

Under settings defined above, our experiments give predictions of cumulative COVID-19 confirmed cases over last 30 days of December 2020. Predictions are shown in Figure 4-11. As we can see, most predictions with Internet information are better than the base case, which is also shown in Table 2. Prediction curves with tweets' information are not as smooth as those without, while Google information leads to a smoother curve closer to the true curve. This might reveal the uncertainty of tweets' information is greater than Google search's.

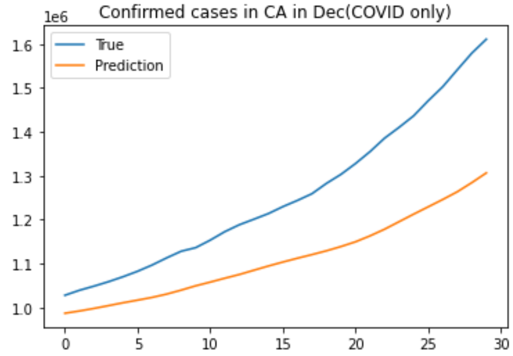

Figure 4 Baseline: COVID-only

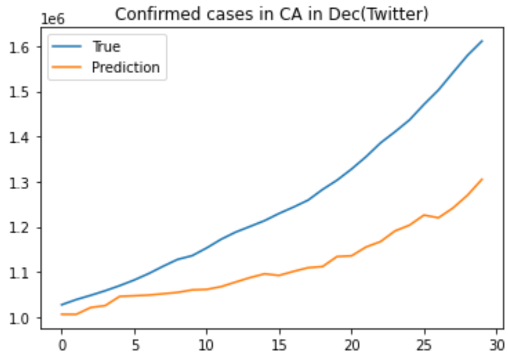

Figure 5 Comparison: Twitter-only

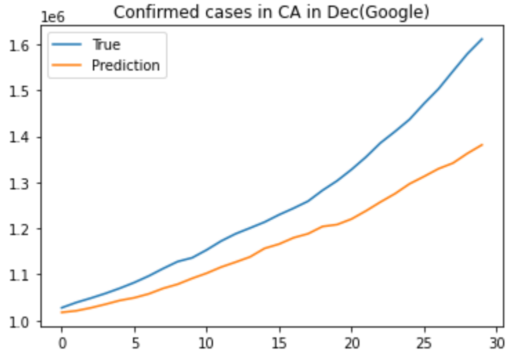

Figure 6 Comparison: Google-only

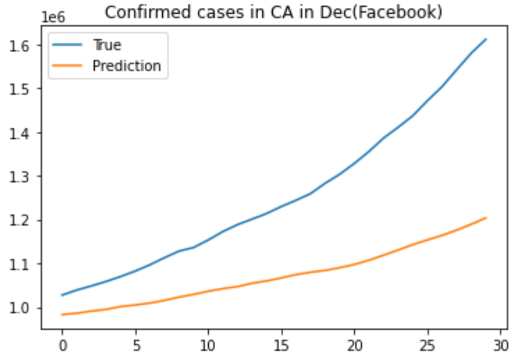

Figure 7 Comparison: Facebook-only

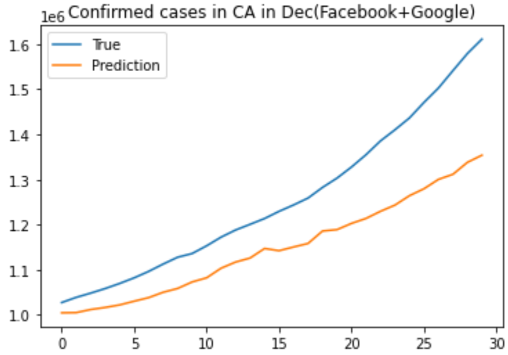

Figure 8 Comparison: Facebook+Google

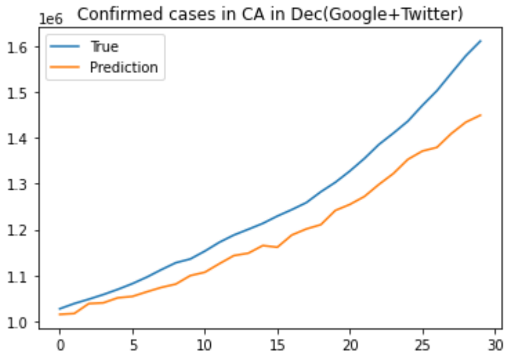

Figure 9 Comparison: Google+Twitter

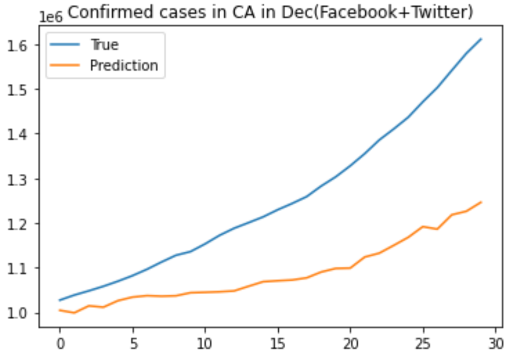

Figure 10 Comparison: Facebook+Twitter

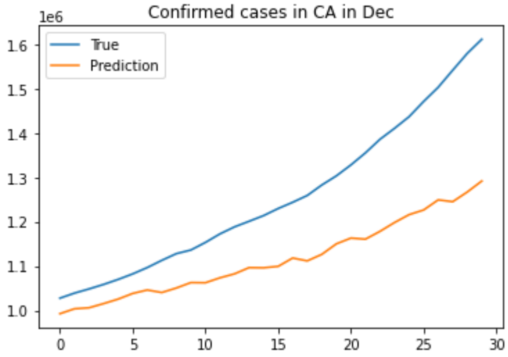

Figure 11 Comparison: All-feature

Another thing that caught our attention is that Internet information included predictions seem to perform much better in short-term prediction but relatively poor in the long term. To be more convincing, we calculated the short-term RMSE and long-term RMSE, shown in Table 2. At short-term predictions, social media informed predictions (e.g., Google RMSE: 0.016) except Facebook-only case, are much better than the baseline (RMSE: 0.043). These might suggest that we should make good use of social media information in short-term predictions.

RMSE of different test periods		
Features	12/02/2020-12/06/2020 (first 5 days)	12/02/2020-12/31/2020
Baseline (COVID	0.154	0.043

only)		
Twitter-only	0.134	0.022
Google-only	0.084	0.016
Facebook-only	0.169	0.046
Twitter+Google	0.059	0.013
Twitter+Facebook	0.154	0.029
Google+Facebook	0.130	0.030
All-feature	0.113	0.026

Table 2 RMSE of experiments in different periods

Besides, we found that combination of Google and Twitter features outperforms all other models. This suggests that it is important for us to properly choose and test features, which might give us a greater model, instead of simply putting powerful features together. To analyze the importance of different features, we applied sensitivity analysis to the ‘Google + Twitter’ experiment, which is our best experiment! Surprisingly, we found that Google features’ influence is much greater than Twitter’s, i.e., 47 : 5.

9.CONCLUSION AND DISCUSSION

1.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.

2.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.

3.LSTM Prediction:

Under most circumstances, Internet-information-informed models perform better than the model with only COVID features, especially in short-term predictions. This might suggest we should take Internet information more serious in short-term prediction. As far as we are concerned, we found Google search features most powerful, and Twitter features useful as well. Moreover, it performs even better when various Internet information sources are combined properly. In a nutshell, let’s not just consider COVID data but more, like Internet information and social media, they help!

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

[14] Delphi Group. (2021). COVIDcast Export Data.DELPHI.https://delphi.cmu.edu/covidcast/export/?sensor=fb-surveysmoothed_wcovid_vaccinated_appointment_or_accept