Navid Shaghaghi*, Andres Calle, George Kouretas, Jaidev Mirchandani, Michael Castillo

eVision: Epidemic Forecasting on COVID-19

Abstract: Vaccination is the primary strategy to prevent COVID-19 illness and hospitalization. However, supplies are scarce and due to the regional mutations of the virus, new vaccines or booster shots will need to be administered potentially regularly. Hence, the prediction of the rate of growth of COVID-19 cases is paramount to ensuring the ample supply of vaccines as well as for local, state, and federal government measures to ensure the availability of hospital beds, supplies, and staff.

eVision is an epidemic forecaster aimed at combining Machine Learning (ML) - in the form of a Long Short-Term Memory (LSTM) Recursive Neural Network (RNN) - and search engine statistics, in order to make accurate predictions about the weekly number of cases for highly communicable diseases. By providing eVision with the relative popularity of carefully selected keywords searched via Google along with the number of positive cases reported from the US Centers for Disease Control and Prevention (CDC) and/or the World Health Organization (WHO) the model can make highly accurate predictions about the trend of the outbreak by learning the relationship between the two trends. Thus, in order to predict the trend of the outbreak in a specific region, eVision is provided with a weekly count of the number of COVID-19 cases in a region along with statistics surrounding common symptom search phrases such as “loss of smell” and “loss of taste” that have been searched on Google in that region since the start of the pandemic. eVision has, for instance, been able to achieve an accuracy of %89 for predicting the trend of the COVID-19 outbreak in the United States.

Keywords: COVID-19 Trend Prediction, Healthcare Technology, Long Short-Term Memory (LSTM) Recursive Neural Network (RNN), Medical Machine Learning (ML)

1 Introduction

Coronavirus Disease 2019 (COVID-19) is the contagious respiratory disease caused by Severe Acute Respiratory

<https://doi.org/10.1515/cdbme-2021-2214>

***Corresponding author: Navid Shaghaghi, Andres Calle, George Kouretas, Jaidev Mirchandani, Michael Castillo:** Ethical, Pragmatic, and Intelligent Computing (EPIC) and BioInnovation and Design (BID) Research Laboratories, Santa Clara University, Santa Clara, California, USA {nshaghaghi, acalle, gkouretas, jmirchandani, mvcastillo}@scu.edu

Syndrome Coronavirus 2 (SARS-CoV-2) which gripped the world in the form of a global pandemic starting in December of 2019 [1] and is ongoing at the time of this writing a year and a half later.

Countless modelling has been conducted with the objective of forecasting the dynamics of transmission and control [2], estimation of the impact of various types of interventions [3][4], or allocation of scarce medical resources [5].

Since most of the forecasting models created to track the COVID-19 pandemic, estimate the number of infected people over time under given conditions based on epidemiological models [6], they have missed the insight discovered by eVision [7] previously on influenza forecasting: by simply providing the ML model with the relative popularity of carefully selected keywords searched via Google along with the number of positive cases reported from the US Centers for Disease Control and Prevention (CDC) and/or the World Health Organization (WHO) the model can make highly accurate predictions about the trend of the outbreak by learning the relationship between the two trends [8].

eVision is an epidemic forecaster aimed at combining ML - in the form of a Long Short-Term Memory (LSTM) Recursive Neural Network (RNN) - and search engine statistics, in order to make accurate predictions about the weekly number of cases for highly communicable diseases [9]. Starting on replicating an older Google model and then improving upon it, predictions are accurately made as far as seven weeks into the future with an accuracy rate of %91 for seasonal influenza [10]. This paper is to report on the adaption and usage of eVision for COVID-19 trend predictions which have yielded an accuracy of %89.

2 Related Work

Lee et al. combine real-time social media data streams with CDC historical datasets using a multilayer perceptron with back propagation in order to predict expected influenza activity for the future, a week or more ahead of time. The keywords analyzed in each tweet include disease types, symptoms, and treatments. The system effectively maps flu activity throughout the US, and was designed to be scalable to other diseases [11].

Bae et al. used an adjusted Susceptible Exposed Infectious Recovered (SEIR) model to portray the effectiveness of information released on social media and used a Monte Carlo method to predict the effect of social media on the spread of COVID-19. They determined that metrics behind social media

use must be included in the adjusted SEIR models to improve the forecast of COVID-19 [12].

Chimmula et al. predicted the outbreak of the COVID-19 disease based on historical transmission data. Recovery rate of COVID-19 patients is an additional parameter used in the model of predicting the outbreak of the disease. A long-short term memory (LSTM) model is used in this work to incorporate the historical transmission data into the prediction method [13].

Chandra et al. incorporate a long-short term memory (LSTM) model, a bi-directional LSTM model, and an encoder-decoder LSTM model to forecast the spread of COVID-19 in select states throughout India. States with the highest infection rates are chosen and compared with states that have experienced an improvement in the infection rates [14].

3 Methodology

The methodology underpinning the application of eVision to COVID-19 is by and large almost identical to its methodology for tracking influenza outbreaks. The data provided to the model are actual case data instead of ILI data along with search trends of illness related keywords, while the underlying LSTM model remained identical [7-10].

3.1 Data Gathering

The data provided to the machine learning model can be divided into two main categories. COVID-19 case data which was gathered and made available to the public by the New York Times [15] and Google keyword search popularity data obtained from Google Trends [16].

3.1.1 COVID-19 Case Data

The COVID-19 case dataset contains various measurements pertaining to the progression of the outbreak. The number utilized for prediction in eVision were those of new daily COVID-19 cases on the national level as well for the states of California, Texas, and New York. These three states were chosen due to their best performance for flu prediction as delineated in [10]. The data was then compressed into weekly results in order to match the scope of the Google Trends dataset.

3.1.2 Google Trends Data

The keywords selected to be pulled from Google Trends focused on COVID-19 specific symptoms. “Loss of Smell” (LOS) and “Loss of Taste” (LOT) were determined to be ideal symptoms of COVID-19 in that they are more specific to the disease than other symptoms. The expected relationship between the keywords and the covid cases nationwide is depicted in Figure 1. A rise in keyword

searches in google generally preceded a rise in case count. Furthermore, the number of searches had some relationship with the increase in cases with the exception of the very first wave of the pandemic where national interest in the disease symptoms was high even in places that had no cases yet. In addition to gathering data about these keywords at a national level, data was also collected on the same keywords for the state of California.

Figure 1: US COVID-19 Cases and Keyword Popularity

3.2 LSTM

An LSTM neural network was easily constructed using functions provided by MATLAB's Deep Learning toolkit. Then a 4x1 layer array was created, consisting of a sequence input layer, an LSTM layer, a fully connected layer, and a regression layer. The LSTM layer was found to work most optimally with 327 hidden units. Its gradient threshold was set to 1, and its learning rate to 0.005. The layer was also assigned a drop factor of 0.1 and given 250 epochs to train over.

3.3 Graphical User Interface

eVision's Graphical User Interface (GUI), depicted in figure 2, was built using MATLAB app designer [2], and provides the user the option to enter unique parameters for personalized predictions. These parameters include the number of weeks they would like to predict ahead (1 or 2 - currently), the prediction level (National or State), and the area they would like to predict for. Preprocessing and data scraping scripts were incorporated into the backend of the interface to allow for real-time predictions based on the parameters entered by the user.

4 Results

Since unlike Influenza case data, COVID-19 data was made available by the CDC and WHO in close to real time, predictions were able to be made at one week in advance. Results were gathered on the accuracy of predictions using no Google Keyword data as well as with all permutations of the

Figure 2: Prototype of user interface for COVID-19

predictions keywords data, as well as keywords data from other states. Predictions were also performed for the state of California alone, in order to determine how well the model would perform on smaller regions.

While SMAPE allows for a quantitative comparison of results between different predictions, certain important metrics are not easily captured by it such as accuracy in predicting the peak week of an outbreak, or the fact that the accuracy of the magnitude of cases is more important during an outbreak than outside of one. As such, qualitative analysis of predictions was conducted alongside quantitative analysis as shown in Table 1.

Scope	Dataset	SMAPE
National	None	0.2282
National	LOS	0.1091
National	LOT	0.1325
National	LOS, LOT	0.1660
National	LOS, LOT, and CA cases and keywords	0.1792
National	CA, NY, TX cases LOS,	0.2404
National	LOT, CA, NY, TX cases and keywords	0.1576
California	None	0.3177
California	LOS	0.3744
California	LOT	0.2956
California	LOS, LOT	0.2559

Table 1: COVID-19 Predictions, One Week Ahead

Across both National and California predictions it can be observed that in almost all cases, predictions are more accurate when made with an augmenting dataset of google keyword trends. This holds with our findings in previous eVision research, showing that the predictive capabilities of the LSTM are greatly enhanced through the acquisition of search term data.

The Californian predictions generally held to the pattern observed in eVision’s influenza forecasting research, where a combination of keywords produces the best predictions

possible. The fact that the accuracy is only 74.41% in the case of using both LOS and LOT keywords, 8.99% less than the equivalent national prediction, is somewhat expected. It has been observed that influenza predictions are more accurate on nation-wide data than specific states. Furthermore, it should also be noted that the predictions constantly underestimated the number of cases in California as seen in Figure 3 – (Right). The national predictions for the United States diverged from results from using eVision for influenza prediction in that rather than the best prediction coming from a combination of national keywords and state level cases and keywords trends, the best prediction came from simply using the “Loss of Smell” keywords achieving an accuracy of 89.09% which performed over three percent better than the next closest prediction, using “Loss of Taste” on its own. The LOS prediction also notably did not have the same issue identified in the California predictions of systematically underestimating COVID-19 cases as can be seen in Figure 4 – (left).

Figure 3: United States LOS Predictions (left) - California LOS and LOT Predictions (Right)

One likely explanation for the differences in performance between the United States prediction and the California predictions is that national google searches will correlate more strongly with national COVID-19 cases. One possible reason for this could be that Californian Google searches could be prompted by national news stories pertaining to outside of the state, which are far more common than the equivalent of international stories that prompt a noticeable increase in searches across the US as a whole.

5 Work in Progress

5.1 Augmenting eVision with Social Media Data

Several work [17][18][19] have utilized social media data in order to perform early or regional predictions of the spread of the COVID-19 pandemic. However, the inclusion of social media data in training eVision for the purpose of enhancing its predictive ability of Influenza outbreak trends proved to have an almost negligible effect. It remains to be seen if the effect of currently being gathered social media data pertaining to COVID-19 symptoms will have a similar effect on eVision’s predictive ability of the COVID-19 outbreak.

6 Conclusion

The prediction of the trend of COVID-19 is paramount to ensuring ample resupplying and application of scarce medical resources. Hence, any forecasting of the trend of regional cases can help medical facilities better prepare and thus succeed in saving more lives. eVision has shown great promise in performing timeseries forecasting of the spread of the infection with an %89 accuracy using CDC or WHO data along with Google Trends data. This methodology has been also successful in predicting the spread of Influenza with %91 accuracy. Hence, this shows that the medical community can benefit greatly from applying machine learning technology to vast amounts of data which may not even have medical nature though which help illuminate the behaviours of the relevant segments of the population.

Acknowledgements

Many thanks are due to Ben Dority, the New-Technology Senior Director at Cepheid Inc. for inspiring the project and supporting it throughout development, as well as to Eva Bouzos, SCU Masters Program Alum and Oncology R&D Research Associate at Cepheid Inc. for her continued support of the project and liaising between the team and Cepheid scientists. Also to Prashanth Asuri, director of SCU's BioInnovation and Design Lab for obtaining financial support from Cepheid Inc. And to Santa Clara University's Frugal Innovation Hub as well as the Departments of Bio Engineering, Computer Science & Engineering, and Mathematics & Computer Science for their continued support of the project.

References

- [1] Page J, Hinshaw D, McKay B (26 February 2021). "In Hunt for Covid-19 Origin, Patient Zero Points to Second Wuhan Market – The man with the first confirmed infection of the new coronavirus told the WHO team that his parents had shopped there". The Wall Street Journal.
- [2] Kucharski AJ, Russell TW, Diamond C, Liu Y, Edmunds J, Funk S, Eggo RM (May 2020). "Early dynamics of transmission and control of COVID-19: a mathematical modelling study". The Lancet. Infectious Diseases. 20 (5): 553–558. doi:10.1016/S1473-3099(20)30144-4. PMC 7158569. PMID 32171059.
- [3] Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, Colaneri M (June 2020). "Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy". Nature Medicine. 26 (6): 855–860. doi:10.1038/s41591-020-0883-7. PMC 7175834. PMID 32322102
- [4] Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N, et al. (May 2020). "The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study". The Lancet. Public Health. 5 (5): e261–e270. doi:10.1016/S2468-2667(20)30073-6. PMC 7158905. PMID 32220655
- [5] Emanuel EJ, Persad G, Upshur R, Thome B, Parker M, Glickman A, et al. (May 2020). "Fair Allocation of Scarce Medical Resources in the Time of Covid-19". The New England Journal of Medicine. 382 (21): 2049–2055. doi:10.1056/NEJMs2005114. PMID 32202722.
- [6] Kermack WO, McKendrick AG (1927). "A contribution to the mathematical theory of epidemics". Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character. 115 (772): 700–721.
- [7] Navid Shaghaghi, Andres Calle, and George Kouretas. 2020. Influenza Forecasting. In *Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS '20)*. Association for Computing Machinery, New York, NY, USA, 339–341. DOI:https://doi.org/10.1145/3378393.3402286
- [8] N. Shaghaghi, A. Calle and Y. Qian, "eVision: Influenza Forecasting Using CDC, WHO, and Google Trends Data," *2020 IEEE / ITU International Conference on Artificial Intelligence for Good (AI4G)*, 2020, pp. 38-45, doi: 10.1109/AI4G50087.2020.9311072.
- [9] N. Shaghaghi, A. Calle and G. Kouretas, "Expanding eVision's Scope of Influenza Forecasting," *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, 2020, pp. 1-10, doi: 10.1109/GHTC46280.2020.9342864.
- [10] Shaghaghi, Navid, et al. "Expanding eVision's Granularity of Influenza Forecasting." *Wireless Mobile Communication and Healthcare: 9th EAI International Conference, MobiHealth 2020, Virtual Event, November 19, 2020, Proceedings 9*. Springer International Publishing, 2021.
- [11] Lee, K., Agrawal, A., & Choudhary, A. (2017, August). Forecasting influenza levels using real-time social media streams. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 409-414). IEEE.
- [12] Bae, Sujin & Sung, Christine & Kwon, Ohbyung. (2021). Accounting for social media effects to improve the accuracy of infection models: Combatting the COVID-19 pandemic and infodemic. *European Journal of Information Systems*. 30. 342-355. 10.1080/0960085X.2021.1890530
- [13] Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals*, 135, 109864
- [14] Chandra, R., Jain, A., & Chauhan, D. S. (2021). Deep learning via LSTM models for COVID-19 infection forecasting in India. *arXiv preprint arXiv:2101.11881*
- [15] New York Times. (2020) Covid-19 Data. <https://github.com/nytimes/covid-19-data>
- [16] Google Inc. (2019). Google flu trends. <https://www.google.org/flutrends/about>
- [17] Qin, Lei, et al. "Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index." *International journal of environmental research and public health* 17.7 (2020): 2365.
- [18] Li, Cuilian, et al. "Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020." *Eurosurveillance* 25.10 (2020): 2000199.
- [19] Xu, Qing, et al. "Characterizing Weibo social media posts from wuhan, china during the early stages of the COVID-19 pandemic: qualitative content analysis." *JMIR Public Health and Surveillance* 6.4 (2020): e24125