Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230290

Digital Disease Surveillance for Emerging Infectious Diseases: An Early Warning System Using the Internet and Social Media Data for COVID-19 Forecasting in Canada

Yang YANG[a,1] Shu-Feng TSAO[a] Mohammad A. BASRI[b] Helen H. CHEN[a] and Zahid A. BUTT[a]
[a]*School of Public Health Sciences, University of Waterloo, Canada*
[b]*Systems Design Engineering, University of Waterloo, Canada*

Abstract. Background: Emerging Infectious Diseases (EID) are a significant threat to population health globally. We aimed to examine the relationship between internet search engine queries and social media data on COVID-19 and determine if they can predict COVID-19 cases in Canada. **Methods:** We analyzed Google Trends (GT) and Twitter data from 1/1/2020 to 3/31/2020 in Canada and used various signal-processing techniques to remove noise from the data. Data on COVID-19 cases was obtained from the COVID-19 Canada Open Data Working Group. We conducted time-lagged cross-correlation analyses and developed the long short-term memory model for forecasting daily COVID-19 cases. **Results:** Among symptom keywords, "cough," "runny nose," and "anosmia" were strong signals with high cross-correlation coefficients >0.8 ($r_{Cough} = 0.825, t-9$; $r_{Runny\ Nose} = 0.816, t-11$; $r_{Anosmia} = 0.812, t-3$), showing that searching for "cough," "runny nose," and "anosmia" on GT correlated with the incidence of COVID-19 and peaked 9, 11, and 3 days earlier than the incidence peak, respectively. For symptoms- and COVID-related Tweet counts, the cross-correlations of Tweet signals and daily cases were $r_{Tweet\ Symptoms} = 0.868, t-11$ and $r_{Tweet\ COVID} = 0.840, t-10$, respectively. The LSTM forecasting model achieved the best performance ($MSE = 124.78, R^2 = 0.88, adjusted\ R^2 = 0.87$) using GT signals with cross-correlation coefficients >0.75. Combining GT and Tweet signals did not improve the model performance. **Conclusion:** Internet search engine queries and social media data can be used as early warning signals for creating a real-time surveillance system for COVID-19 forecasting, but challenges remain in modelling.

Keywords. Emerging infectious diseases, COVID-19, digital surveillance systems, internet search engines, social media, Google Trends, Twitter, long short-term memory (LSTM)

1. Introduction

Emerging infectious diseases (EIDs), including COVID-19, Ebola virus, and SARS, pose significant threats to global public health and economies. As of March 1, 2023, the global

[1] Corresponding Author: Yang Yang, E-mail: y24yang@uwaterloo.ca.

confirmed cases of COVID-19 stand at approximately 675.4 million, with roughly 6.8 million deaths reported [1]. The economic costs of EIDs, such as COVID-19, are also staggering. According to the Canadian Institute for Health Information (CIHI), in 2021, due to the impact of the COVID-19 pandemic, Canada's healthcare costs were expected to reach an all-time high of $308 billion. This translates to roughly $580,000 per minute, with healthcare expenses accumulating to around $70,000 in the time it takes to read this sentence, or approximately $10,000 per second [2]. During global epidemics, the general public tends to utilize media, such as internet searches and social media, for health information more frequently than ever. This heightened consumption of online and social media data can facilitate real-time surveillance of emerging diseases and the prediction of epidemics [3-4]. Our team conducted a scoping review on COVID-19, which identified a gap in research and implementation of a real-time surveillance system for COVID-19 [5]. An efficient real-time surveillance system would serve as an early warning system, empowering public health authorities and hospitals to respond quickly to EID threats. Google Trends (GT) and Twitter can be used for digital disease surveillance in real-time, such as the HealthMap system [6], which uses machine learning algorithms to track outbreaks, and the Pandemic Response Platform [7] developed by the WHO and Microsoft to track disease spread and provide real-time updates to public health officials. This study aims to evaluate the association between internet search engine queries such as GT and social media data on Twitter about COVID-19 in Canada and investigate if information from these sources has predictive power as early warning signals to predict COVID-19 cases.

2. Methods

2.1. Data Collection and Preprocessing

We extracted GT and Twitter data from 1/1/2020 to 3/31/2020 in Canada as early warning signals, including symptom keywords (‘cough,’ ‘runny nose,’ ‘anosmia,’ ‘sore throat,’ ‘shortness of breath,’ ‘fever,’ ‘headache,’ ‘body ache,’ ‘dyspnea,’ and ‘fatigue’) from GT and COVID-19-related hashtags (‘pneumonia,’ ‘cough,’ ‘fever,’ ‘running nose,’ and ‘breath’) from Twitter, respectively. COVID-19 data was obtained from the COVID-19 Canada Open Data Working Group. All data were normalized to the same scale for analysis.

2.2. Data Analysis

We applied Fast Fourier Transform (FFT), Moving Average (MA), Savitzky–Golay (SG), and Lowess smoothing methods to remove the white noise obtained in the data. Denoising was done using adjacent averaging on both GT and Tweet signals. We performed time-lagged cross-correlation analyses using denoised signals to examine the relationship between each signal and daily COVID-19 cases. We then developed the long short-term memory (LSTM) model for forecasting COVID-19 cases with TensorFlow and Keras in Python. We used the past 20 days’ data for model training. The LSTM models were trained using the Adam optimizer, and the hidden units were 148 through experiments. We calculated the loss function, mean squared errors (MSE), and R^2 and adjusted R^2 to evaluate the model performance. We further fine-tuned the initial learning rate and learning rate drop hyperparameters, ranging between 0.0001 and 0.001.

3. Results

3.1. Time-Lagged Cross-Correlation Analysis

The time-lagged cross-correlation analysis between GT signals and daily COVID-19 cases showed a range of [1, 13] days of time lag. The maximum correlation coefficients varied greatly, ranging between -0.275 (Fatigue) and 0.825 (Cough). The cross-correlation analysis between Tweet signals and daily COVID-19 cases showed similar lags and coefficients. Table 1 summarizes the cross-correlation analysis results. Figure 1 shows the implementation of denoising the GT data using the MA method (left) and the denoised GT and Tweet signals (right).

Table 1. Time-lagged cross-correlation analyses on Google Trends and Twitter data.

Signals	Lag (in days from time t)	Max Correlation Coefficient
	Google Trends (GT)	
Cough	-9	0.825
Runny Nose	-11	0.816
Anosmia	-3	0.812
Sore Throat	-6	0.790
Shortness of Breath	-9	0.762
Fever	-10	0.752
Headache	-8	0.723
Body Ache	-5	0.612
Dyspnea	-13	0.501
Fatigue	-1	-0.275
	Tweet	
Symptom related	-11	0.868
COVID-19 related	-10	0.840

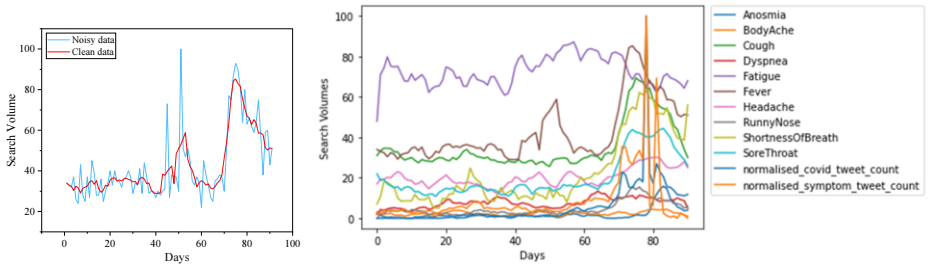

Figure 1. Signal processing. **Left**: Noisy and clean data for Google Trends for the keyword 'Fever.' **Right**: Denoised Google Trends and Tweet signals from 1/1/2020 to 3/31/2020 in Canada.

3.2. LSTM Forecasting Models

We experimented with different subsets of features using GT, Tweet, and the combination of GT and Tweet signals for LSTM forecasting models. Using GT signals with a correlation coefficient >0.75 achieved the best model performance ($MSE = 124.78, R^2 = 0.88, adjusted\ R^2 = 0.87$). Table 2 compares LSTM modelling results.

Table 2. Long short-term memory (LSTM) modelling results. **M1**: 'Cough,' 'Runny Nose,' 'Anosmia,' 'Sore Throat,' 'Shortness of Breath,' 'Fever,' 'Headache,' 'Body Ache.' **M2**: 'Cough,' 'Runny Nose,' 'Anosmia,' 'Sore Throat,' 'Shortness of Breath,' 'Fever,' 'Headache.' **M3**: 'Cough,' 'Runny Nose,' 'Anosmia,' 'Sore Throat,' 'Shortness of Breath,' 'Fever.' **M4**: 'Cough,' 'Runny Nose,' 'Anosmia.' **M5**: Tweet symptoms-related counts. **M6**: Tweet symptoms-related counts, Tweet COVID-related counts. **M7**: M3+M5. **M8**: M3+M6. **MSE**: mean squared error. **Adj R^2**: adjusted R^2.

Model	M1	M2	M3	M4	M5	M6	M7	M8
MSE	413.40	495.35	124.78	496.62	23500.5	26611.8	325.57	365.42
R^2	0.59	0.51	0.88	0.51	-22.89	-26.06	0.68	0.64
Adj R^2	0.57	0.48	0.87	0.48	-23.81	-27.10	0.66	0.62

Figure 2 illustrates the performance of the best model, M3.

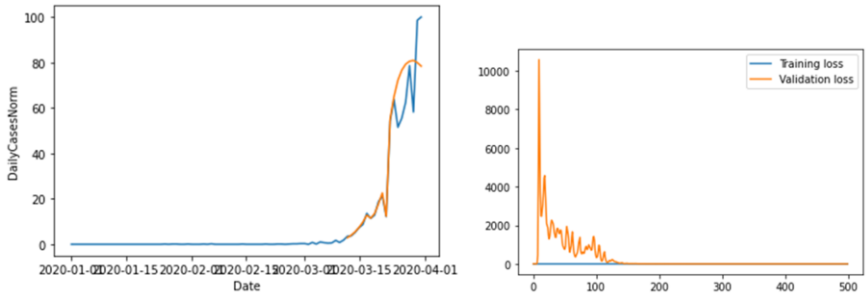

Figure 2. The long short-term memory (LSTM) model uses Google Trends signals with a maximum correlation coefficient >0.75. **Left**: COVID-19 confirmed cases in Canada (blue line: true values; orange line: predicted values). **Right**: training and validation loss (x-axis: epochs; y-axis: loss as measured by mean squared errors)

4. Discussion

This study uses denoised internet search engine queries (GT) and social media (Twitter) data to predict COVID-19 cases in Canada, demonstrating that these sources can be used for digital disease surveillance as an early warning system for EIDs. The choice of denoising method depends on the characteristics of the data and desired accuracy and reliability. FFT was found to be difficult to use on our data due to the lack of character frequencies in the white noise. MA, SG, and Lowess worked better than FFT and are suitable for data with a consistent trend, though they may need to be more effective for data with sudden changes in trend. GT data had more noise than Tweet data, possibly due to the normalized nature of GT data.

Using time-lagged cross-correlation analysis, this study found that COVID-19 symptom-related search terms strongly correlated with daily COVID-19 cases with a time lag of 1-13 days. GT and Twitter data can serve as early warning signals for real-time digital disease surveillance. Still, the LSTM model using GT data performed better than the model using Twitter data or a combination of both. Although tweets can be used as early warning signals for COVID-19 outbreaks, they are reactive and less effective than GT in predicting cases. According to a study, a correlation has been observed between such data and the incidence of COVID-19, with the data peaking 10-14 days before the incidence peak [8]. Other social media platforms like Facebook [9] and Reddit [10] can also provide valuable data for real-time disease surveillance for EIDs. Our study highlights the potential of internet search engines and social media data for real-time

disease surveillance. Still, challenges, such as limited data at the beginning of the COVID-19 pandemic, changing conditions, and unforeseen events, can impact accuracy. In addition, social media data may be subject to biases, such as the under-representation of certain groups, and may not always capture the full picture of an EID outbreak.

Strengths of this study include utilizing signal processing techniques and LSTM modelling to analyze internet search queries and social media data. However, GT data has a lower resolution and only provides relative search volume, and this study only used English Tweets, which may exclude valuable information. Additionally, there may be issues with the accuracy and reliability of Twitter's geolocation. Furthermore, retrospective time-series analyses are useful for monitoring disease outbreaks but can be prone to over-fitting, especially for complex data with many variables. Internet and social media data can provide valuable insights but should be considered alongside other surveillance sources to ensure a comprehensive understanding of disease trends. A multi-faceted strategy, such as incorporating multiple data sources and multimodal modelling with infectious disease models, is warranted for accurate and comprehensive EID surveillance. Lastly, identifying relevant symptom keywords of an EID requires a dynamic and adaptive approach that integrates various information sources and is continually refined as new data becomes available. Our future research will employ an ontology-based method to systematically identify and organize pertinent symptom keywords for an EID, even before they are commonly recognized or reported.

5. Conclusions

A real-time digital disease surveillance system that utilizes internet search engine queries and social media data can be an early warning system for forecasting EIDs like COVID-19. Such a system has the potential to assist in epidemiological control and monitor public perceptions of the disease, as well as forecast trends in outbreaks. However, challenges in modelling arise due to the noise of self-generated data and the identification of relevant symptom keywords of an EID before they are publicly known.

References

[1] Johns Hopkins University and Medicine. Coronavirus resource center: world map. https://coronavirus.jhu.edu/map.html. Date accessed: January 5, 2023
[2] COVID-19 expected to push Canada's health spending beyond $300 billion in 2021. https://www.cihi.ca/en/news/covid-19-expected-to-push-canadas-health-spending-beyond-300-billion-in-2021 Date accessed: December 8, 2022
[3] Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. Journal of medical Internet research. 2020 Apr 21;22(4):e19016.
[4] Al-Garadi MA, Khan MS, Varathan KD, Mujtaba G, Al-Kabsi AM. Using online social networks to track a pandemic: A systematic review. Journal of biomedical informatics. 2016 Aug 1;62:1-1.
[5] Tsao SF, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. The Lancet Digital Health. 2021 Mar 1;3(3):e175-94.
[6] About HealthMap https://healthmap.org/about/. Date accessed: December 8, 2022
[7] Pandemic Prevention Platform (P3) https://www.darpa.mil/program/pandemic-prevention-platform Date accessed: December 8, 2022
[8] Li C, et al. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Eurosurveillance. 2020 Mar 12;25(10):2000199.
[9] Hossain L, Kam D, Kong F, Wigand RT, Bossomaier T. Social media in Ebola outbreak. Epidemiology & Infection. 2016 Jul;144(10):2136-43.
[10] Hu M, Conway M. Perspectives of the COVID-19 Pandemic on Reddit: Comparative Natural Language Processing Study of the United States, the United Kingdom, Canada, and Australia. JMIR infodemiology. 2022 Sep 27;2(2):e36941.