Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Pattern Recognition Letters 158 (2022) 133–140

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

An infodemiological framework for tracking the spread of SARS-CoV-2 using integrated public data

Zhimin Liu, Zuodong Jiang, Geoffrey Kip, Kirti Snigdha, Jennings Xu, Xiaoying Wu, Najat Khan, Timothy Schultz*

Janssen R&D Data Science, Janssen Research and Development, 2341 S Whittmore St, Titusville 08560, Furlong, PA 18925, United States

ARTICLE INFO

Article history:
Received 20 April 2021
Revised 11 February 2022
Accepted 22 April 2022
Available online 26 April 2022

Edited by Maria De Marsico

Keywords:
Infodemiology
News mining
Word2Vec
Signal burst model
Google trends
Prophet model

ABSTRACT

The outbreak of the SARS-CoV-2 novel coronavirus has caused a health crisis of immeasurable magnitude. Signals from heterogeneous public data sources could serve as early predictors for infection waves of the pandemic, particularly in its early phases, when infection data was scarce. In this article, we characterize temporal pandemic indicators by leveraging an integrated set of public data and apply them to a Prophet model to predict COVID-19 trends. An effective natural language processing pipeline was first built to extract time-series signals of specific articles from a news corpus. Bursts of these temporal signals were further identified with Kleinberg's burst detection algorithm. Across different US states, correlations for Google Trends of COVID-19 related terms, COVID-19 news volume, and publicly available wastewater SARS-CoV-2 measurements with weekly COVID-19 case numbers were generally high with lags ranging from 0 to 3 weeks, indicating them as strong predictors of viral spread. Incorporating time-series signals of these effective predictors significantly improved the performance of the Prophet model, which was able to predict the COVID-19 case numbers between one and two weeks with average mean absolute error rates of 0.38 and 0.46 respectively across different states

1. Introduction

COVID-19, the disease caused by the SARS-CoV-2, has been rapidly spreading across the globe and has become a substantial health threat worldwide. As of March 20, 2021, an estimated 122 million people worldwide have been infected with the virus, with an estimated 2.7 million deaths [1]. Accurately monitoring and forecasting regional progression of COVID-19 can help (1) healthcare systems to ensure sufficient supply of equipment and personnel to reduce fatalities, (2) the pharmaceutical industry to perform clinical trials for vaccines or medicines, and (3) world governments to make or adjust non-pharmaceutical interventions or vaccination plans.

Internet sources and data have been employed to inform public health and policies; this application is referred to as Infodemiology (i.e., information epidemiology) [2]. These data sources have in the past nowcasted and forecasted outbreaks and epidemics of various infectious diseases [3–8]. During a pandemic, leveraging infodemiological data, especially in the early phase of the pandemic when there is not enough infection data to generate accurate models, can be a practical way to monitor viral transmission and help the governments to take action more quickly.

Of particular interest to infodemiology as applied to COVID-19 is news media, which serves as a crucial communication medium that can significantly affect individuals' behavior. News media data can be used to study the sentiment of the society in response to COVID-19-related policies and vaccinations [9]. Media coverage on these topics and the corresponding sentiments can also potentially be useful predictive factors for COVID-19 cases. To capture specific news in unstructured formats, Natural Language Processing (NLP) techniques are required; however, commonly used topic modeling methods like Latent Dirichlet Allocation (LDA) [10] have poor performance when analyzing COVID-19-related news as articles tend to repeat very similar vocabularies. This makes it difficult to parse specific subjects (e.g., COVID-19-related school reopening vs. lockdown). Thus, a new NLP method is needed.

Google Trends (GT) is another popular infodemiology data source that is actively used in health and medicine to track and forecast diseases and epidemics [11]. Several papers have used GT data to monitor, track, and forecast COVID-19 in the US [12–14]. These studies consistently identified a high correlation between

* Corresponding author.
E-mail address: tschult4@its.jnj.com (T. Schultz).

https://doi.org/10.1016/j.patrec.2022.04.030

GTs of COVID-19 related terms and new COVID-19 cases for a lag period ranging from 12 to 16 days, demonstrating the strong predictive power of GT for COVID-19 progression. However, most of these studies were conducted in the early stages of the pandemic; at this time, the pandemic has lasted for more than a year, with many regions in US having experienced at least two waves. People's behaviors, such as online search activities, may change as the pandemic evolves. For example, familiarity with COVID-19-related information increases since the beginning of the pandemic and therefore certain search terms may fall out of interest. Therefore, it is necessary to re-examine the leader-follower relationship between GT and COVID-19 case numbers with the more recent and comprehensive data.

An emerging data source to track the spread of SARS-CoV-2 comes from wastewater monitoring [15–17,]. Monitoring sewage for viral RNA concentrations enables effective population-level surveillance, providing a sensitive signal of its circulation throughout communities. This data unbiasedly captures circulation without the need for conducting PCR testing or unaccounted asymptomatic cases. It has been shown that viral concentrations of wastewater are 0-10 days ahead of clinically diagnosed new COVID-19 cases [15,17], suggesting another predictor to forecast COVID-19 cases. As the US government initiates the National Wastewater Surveillance System in response to the COVID-19 pandemic, more data from different regions will be collected and reported.

While several models have been used to forecast COVID-19 cases, many of them cannot integrate multiple time series signals. Auto Regressive Integrated Moving Average (ARIMA) model has been used by researchers around the world to forecast the spread of this pandemic and generates accurate predictions [18–20]. ARIMA works best when data is stationary, meaning that the variance and the mean of the data remain constant over time. In addition, ARIMA can only be implemented on univariate time series. Kalman filtering is another algorithm used to forecast COVID-19 cases but only produces satisfying short term (daily) predictions [21,22]. Prophet has been widely used and accepted due to its accuracy and ease of usability [23,24]. Its automatic nature gives flexibility to time series data that have dramatic changes so that users do not have to worry about their data being not suited for the model [23]. More importantly, Prophet provides the option to integrate other time-series covariates and thus serves as an ideal model to combine various digital data streams.

2. Our contribution

In this study, we extracted informative signals from three available public datasets (i.e., news websites, Google Trends, and wastewater SARS-COV-2 measurements), and for the first time integrated these predictive signals into a model to forecast COVID-19 trends. We first built an effective NLP pipeline using the Word2Vec embeddings from pre-trained deep neural network [25] on Google News to identify news on specific topics. We validated the pipeline by successfully distinguishing different groups of news and identifying "school reopen" and "lockdown" related COVID-19 news. We further identified the time points when specific news broke out abruptly using a burst detection model. We then aligned various signals with new COVID-19 cases, checked their correlations and synchronies, and identified several signals as early indicators of enhanced spread. Finally, we integrated the selected signals into a Prophet model to predict future COVID-19 cases and demonstrated that these signals could significantly improve the base model's performance.

3. Experimental

3.1. Data sets

COVID-19 case numbers The number of daily cumulative confirmed cases in 19 states of United States were obtained for the period until Dec 31, 2020 from the COVID Tracking project (https://covidtracking.com/data/download). The column of "positive" that contains total number of confirmed plus probable cases of COVID-19 reported by the state was used. For Massachusetts data, there is one time point that has a smaller cumulative case number than that of the previous week. That week's case number was replaced with the average case number within that month.

News articles The public news data in this study were obtained from NewsAPI.org, which allows to search public news and articles from over 30,000 news sources in 54 countries, including ABC, BBC, Australian Financial Reviews and others. COVID related news in each state were acquired by keywords searching of COVID terms AND a specific state's name in each article's title. COVID terms include "coronavirus", "COVID-19", "COVID19", "SARS-CoV-2", "2019-nCov". Eventually, 33,083 relevant news and articles were collected from 19 states with a period of Dec. 1, 2019 to Dec. 31, 2020.

Google trends Google Trends provides the relative search volume for each keyword. This value is calculated by dividing the total number of searches for a keyword by the total searches within a geographic and time range. Keywords can be filtered by location with a resolution from worldwide to a specific city and time span. Time series data are presented on a normalized scale of 0 to 100, where 0 represents no search and 100 represent peak search activity for a particular keyword or string. Google Trends' daily base data were mined in this study from February 1, 2020, to December 31, 2020. The following keywords were searched: COVID symptoms, COVID testing, covid rapid testing, school opening, lockdown. Data for each keyword with in each of 19 selected states were obtained.

Wastewater COVID-19 measurement Wastewater COVID-19 data in Massachusetts (MA), Ohio (OH), and Arizona (AZ) were obtained through the following links: MA: https://www.mwra.com/biobot/biobotdata.htm; OH: https://coronavirus.ohio.gov/wps/portal/gov/covid-19/dashboards/other-resources/wastewater; AZ: https://data.tempe.gov/datasets/covid-wastewater-results-public-view/data?selectedAttribute=Day. Data in Arizona was limited to Tempe City, while MA and OH data came from multiple sites across the states. MA had the most consistent daily measurements since March 2020, while the measurements in OH and AZ were less frequent since July and April in 2020. The units of all the data were converted into number of copies per liter of wastewater. Data of MA and AZ were averaged by week, while in OH data were aggregated weekly by taking the median measurements across multiple sites and different days to the reduce the effect of some extreme measurements.

3.2. Analytical methods

Sentiment The sentiment score for each news was calculated by aggregating the polarity scores from each word in the title using the vaderSentiment package from NLTK.

Search news of a specific topic The title, headline, and content of each news item were concatenated and all characters were converted to lower case. Punctuation, text in square brackets, words containing numbers, stop words from NLTK, and state names were removed. The remaining words were tokenized and lemmatized with NLTK packages. The Word2Vec embeddings from pre-trained deep neural network on Google News (https://s3.amazonaws.com/dl4j-distribution/GoogleNews-vectors-negative300.bin.gz) were imported and used to vectorize each

news item with the mean of word embeddings for each word in that article. Similarly, specific topics represented by some keywords were vectorized by the mean word embeddings for each word (e.g., "school reopen": "school reopen reopening schools operating schools students teachers"; "lockdown": "lockdown restrict restrictions").

The similarity between each news item and a specific topic was measured by calculating the cosine of the angle between the two vectors that represent them:

$$\cos(\theta) = \frac{A \cdot B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\ \sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A and B represent the vectors of the news and the topic, respectively.

Cosine similarity scores of 1 and –1 represent two overlapping vectors and two exactly opposite vectors, respectively. The distributions of cosine similarity scores between each news item and the target topic were examined and a threshold of mean + 2 × standard deviations was chosen to qualitatively identify news related to the target topic.

Cluster news The sum of squared distances of samples to their closest cluster center and the mean Silhouette Coefficient of all samples were calculated with functions in scikit-learn package: K-Means (default settings) and silhouette_score (metric = "cosine") when vectorized news item were split into different numbers of clusters (1–50) using K-Means. The optimal number of clusters leading to a low intra-cluster distance and a high mean Silhouette Coefficient was used. The most frequent words within each cluster were used to represent each cluster.

Aggregate data by week Daily COVID-19 news numbers were summed by week since the daily news numbers were generally small (< 10) in most states. The fraction of news related to a specific topic was calculated by dividing the number of target news items by the total number of news items in each week for each state. New weekly COVID-19 cases were derived by differencing accumulative case numbers reported in each state. The weekly average Google Trends of specific terms and wastewater COVID-19 measurement were used.

Burst model The "burst_detection" package (https://github.com/nmarinsek/burst_detection) that implements Kleinberg's burst detection algorithm for batched data [26] was used in this study. In this model, there are two possible states: baseline state (lower probability) and bursty state (higher probability). The probability of baseline state (p_0) is the overall proportion of target events:

$$p_0 = \frac{R}{D}$$

where R is the sum of daily target news items (e.g., "school reopen" or "lockdown" related news or news with negative sentiments) and D is the sum of all daily news items in each week.

The bursty state probability (p_1) is equal to the baseline probability multiplied by some constant s.

$$p_1 = s \times p_0$$

Based on the news data, $s = 2$ was used to detect bursts for "school reopen" or "lockdown" related news and $s = 1.2$ was used for COVID-19-related news with negative sentiments.

Two things could determine which state the system is in at any given time: the difference between the observed proportion and the expected probability of each state denoted by sigma:

$$\sigma(\mathrm{i}, \mathrm{r_t}, \mathrm{d_t}) = -\ln\left[\binom{d_t}{r_t} p_i^{r_t} (1 - p_i)^{d_t - r_t}\right]$$

where i is the state (0: baseline state; 1: bursty state), d_t and r_t are the number of target news and total news in each week. the difficulty of transitioning from the previous state to the next state denoted by the transition cost, tau:

$$\tau = (i_{next} - i_{prev}) \times \gamma \times \ln(n)$$

where n is the number of weeks and γ is the difficulty of transitioning to higher states. Note that there is no cost associated with staying in the same state or returning to a lower state. γ is critical to exclude false bursts generated from a small number of news, and thus a specific γ was chosen to ensure that only time points with enough news (more than the median of total news number across all time points) were identified as bursts.

The total cost of transitioning from one state to the other is equal to the sum of two functions above. The optimal state sequence q that minimized the total cost would be characterized by the Viterbi algorithm. The *weight* of a burst that begins at *t1* and ends at *t2* can be estimated with the following function:

$$weight = \sum_{t=t1}^{t2} (\sigma(0,\ r_t,\ d_t) - \sigma(1,\ r_t,\ d_t))$$

Burst weight demonstrates how much cost is reduced when the system is in a burst state versus the baseline state during the burst period. The greater the weight, the stronger the burst would be.

Correlation analysis Spearman's rank correlation coefficient was calculated to determine the correlation between weekly COVID-19 case numbers and various weekly aggregated signals. Time-lagged cross-correlation (TLCC) was used to identify the directionality between two time-series signals such as a leader-follower relationship in which the leader (e.g., Google Trends and news volume) initiates a response which is repeated by the follower (COVID-19 case numbers). TLCC was measured by incrementally shifting one time series vector and repeatedly calculating the correlation coefficient between two signals. Specifically, it was implemented with pandas functionality (datax.corr(datay.shift(lag)), where datax and datay are two time series signals and lag is the shifting window). The correlation analyses were made for each state individually.

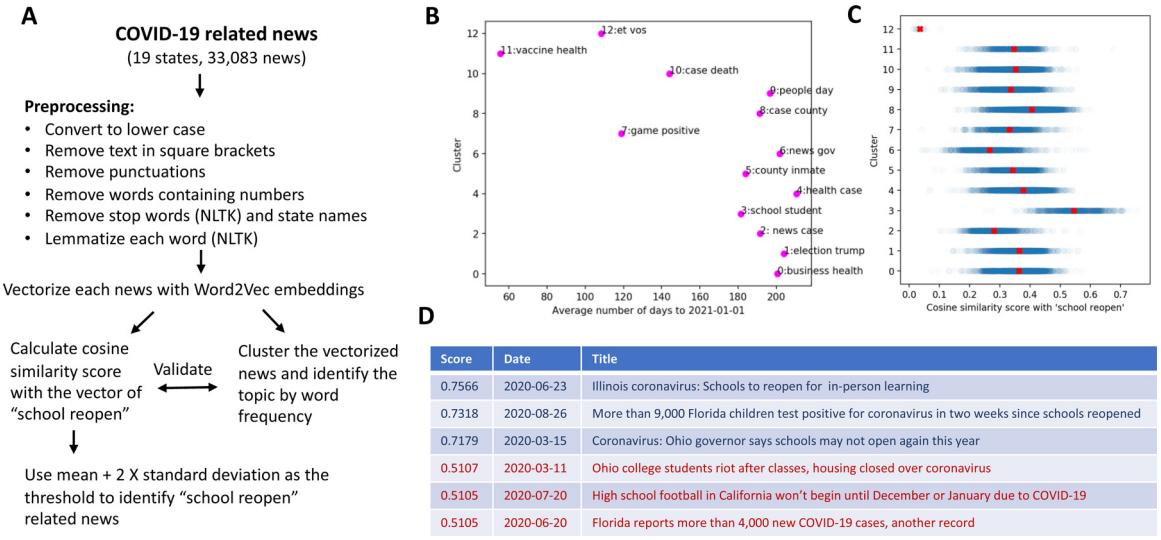
Predict COVID-19 trends with Prophet model In this study due to the nature of weekly aggregated data, the component of daily/weekly/yearly seasonality or holiday was not used in the Prophet model [23]. Only time-series signals that were shown as early predictors of COVID-19 trends (Google Trends of "COVID testing", "covid rapid testing", "COVID symptoms", and "lockdown", and COVID-19 related news volume: "news count") were shifted and used as extra regressors in the Prophet model via the "add_regressor" function. The extra regressor must be known for both historical and future dates. Therefore, it must either be something that has known future values or something that can be forecasted elsewhere. Here, the time series data of these early predictors were shifted by one or two week(s) to generate the future values since they were leading COVID-19 case numbers for 1-3 weeks from the correlation analyses. Columns with these extra regressor values were put into both the fitting and prediction data frames. The default settings were used in the Prophet model since performance was not improved by tuning parameters like "mode" and "prior_scale". The Prophet model was applied to data from each state. The data from the last five weeks were used as the test data to measure its performance. The following statistical measures were used:

Mean Absolute Error (MAE):

$$MAE = \frac{1}{N} \sum_{k=1}^{N} |z_k - \widehat{z_k}|$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{N} \sum_{k=1}^{N} \left| \frac{z_k - \widehat{z_k}}{z_k} \right|$$



Score	Date	Title
0.7566	2020-06-23	Illinois coronavirus: Schools to reopen for in-person learning
0.7318	2020-08-26	More than 9,000 Florida children test positive for coronavirus in two weeks since schools reopened
0.7179	2020-03-15	Coronavirus: Ohio governor says schools may not open again this year
0.5107	2020-03-11	Ohio college students riot after classes, housing closed over coronavirus
0.5105	2020-07-20	High school football in California won't begin until December or January due to COVID-19
0.5105	2020-06-20	Florida reports more than 4,000 new COVID-19 cases, another record

Fig. 1. Search for "school reopen" related news with an NLP pipeline using Word2Vec embeddings. (A) Detailed procedures to identify "school reopen" related COVD-19 news. (B) The times of occurrence of 13 clusters of news, each of which is represented by its two most frequent words. (C) The distribution of cosine similarity scores with "school reopen" in each cluster, red asterisks represent the mean. (D) Examples of "school reopen" related COVID-19 news that pass the threshold (0.5104). Blue and red mark news that have the largest and smallest cosine similarity scores, respectively.

Table 1
Notations used in this article.

Parameters	Notation Used
Cosine similarity score	$\cos(\theta)$
Vector representing each news	A
Vector representing the target topic	B
Probability of baseline state	p_0
Probability of bursty state	p_1
Sum of daily target news numbers	R
Sum of all daily news numbers	D
Constant to calculate the bursty state probability	s
Difference between observation and expectation	σ
State	i
Week	t
Number of target news of the week	r_t
Number of all news of the week	d_t
Transition cost	τ
Difficulty of transitioning to higher states	γ
Number of weeks	n
Optimal state sequence	q
Burst weight	$weight$
Mean absolute error	MAE
Mean absolute percentage error	$MAPE$
Total number of weeks in the test data	N
Week in the test data	k
Actual value of the week	z_k
Predicted value of the week	$\widehat{z_k}$

where z_k denotes the actual value and $\widehat{z_k}$ denotes predicted value for the k_{th} week. N is the total number of weeks in the test data (Table 1).

4. Results

4.1. Identify news of a specific topic

The NLP pipeline to extract news on a specific topic like "school reopen" was shown in Fig. 1A. Each news and target topic were first preprocessed and then vectorized by averaging word embeddings in the article and topic, respectively. A cosine similarity score was calculated using these two vectors. Thresholding the cosine similarity scores could identify news related to the target topic. The identified news as well as clustering news based on the vectors were manually examined to validate the pipeline. 33,083 vectorized COVID-19 news from 19 states were divided into 13 clusters based on K-means clustering algorithm (Fig. S1) and each cluster was represented by the most frequently occurring words (Fig. 1B and Table S1). Different groups of news like "election" (cluster 2), "school" (cluster 3), "prison" (cluster 5), "sports" (cluster 7), and "vaccine" (cluster 11) were successfully characterized by clustering their vectors. Furthermore, the times of occurrence of these clusters were consistent with real situations as shown in Fig. 1B. For example, "school"-related news (cluster 3) occurred in June 2020 when the second wave of COVID-19 arrived in US, leading to numerous discussions on school opening/closing. The occurrence of "sports" news (cluster 7) was centered in September when major sports leagues started or resumed their seasons (NFL: 09/10/2020; NBA: 07/30/2020; MLB: 07/23/2020). Most of "vaccine" related news was reported in the end of the year when the first two COVID-19 vaccines were approved for emergency use (Pfizer-BioNTech: 12/11/2020; Moderna: 12/18/2020). These results demonstrated vectors generated from Word2Vec embeddings accurately captured the information in the news.

To identify the news of a specific topic, a cosine similarity score was calculated between the vectors representing the topic (e.g., "school reopen") and the news. As shown in Fig. 1C, news in cluster 3 representing "school" had significantly high cosine similarity scores with "school reopen" compared to other clusters. The distributions of cosine similarity scores of all news articles with "school reopen" and "lockdown" were shown in Fig. S2A and S2B, respectively. An arbitrary threshold of mean + 2 × SD shown by the red dashed lines was used to identify the target news items. Titles of some examples related to "school reopen" and "lockdown" were shown in Fig. 1D and Fig. S2C, respectively. While news with higher similarity scores were more related to the topic, most selected news items that passed the threshold was associated with the chosen topic. Therefore, this pipeline is very efficient to sort out news based on search terms. It is noteworthy that search terms can be tweaked to obtain a more accurate vector to represent a specific topic.

4.2. Correlation between various signals and COVID-19 case numbers

Besides "school reopen"- and "lockdown"-related news, many other signals were extracted from news, GT data, and viral mea-

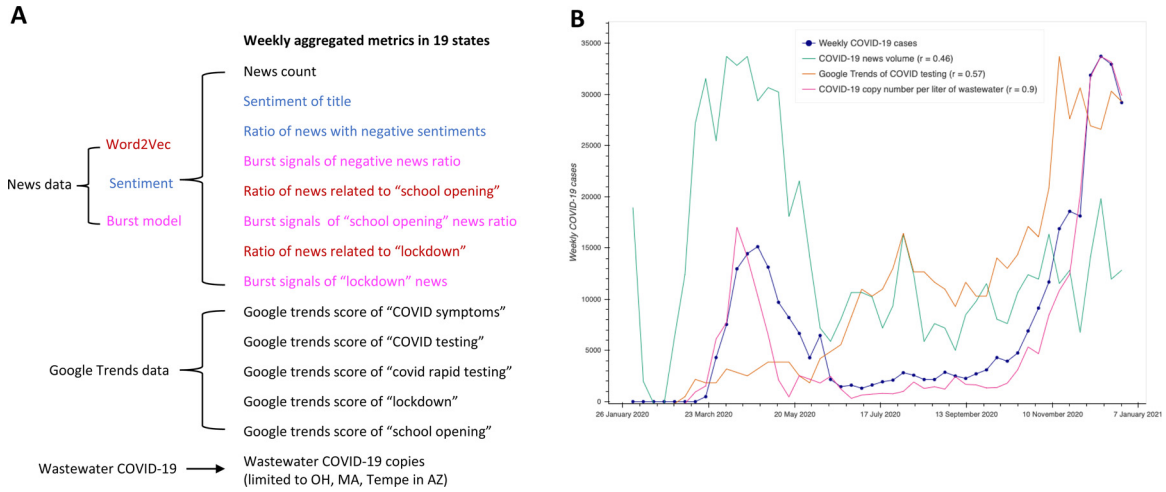


Fig. 2. Align various signals with COVID-19 case numbers. (A) signals extracted from various data sources to track COVID-19 cases. (B) Several signals correlate well with COVID-19 case numbers in Massachusetts. r represents the Spearman's rank correlation coefficient.

surements in wastewater as shown in Fig. 2A. Since the signals extracted from news were generally related to government policies, pivotal events, or public opinions which could influence COVID-19 trends, Kleinberg's burst detection model [26] was used to detect "bursts of activity" when these signals increased sharply (see Burst model in Experimental), aiding in the monitoring of epidemic spread (Fig. S3). All signals were then aligned with COVID-19 case numbers and their correlations were examined in each state (Figs. 2B and S4). Across different states, GT of COVID-19 related terms (e.g., "COVID testing", "covid rapid testing", "COVID symptoms") and wastewater COVID-19 measurement correlated well with COVID-19 cases (Figs. 2B and S4), while signals obtained from news and their bursts had variable correlations with COVID-19 cases (Fig. S3 and S4). In some states like Massachusetts and Arizona, correlation of news volume with COVID-19 cases were high (Figs. 2B and S4). In addition, many redundant signals correlated well with each other as shown in Fig. S5. It is noteworthy the counts of various specific news generally correlated well with total news count, indicating count signals of news of various topics were biased by total number of news items and thus the ratio signals that were normalized by total count of news items were used in the following analyses.

4.3. Synchrony analysis

To further identify signals that could be predictors of COVID-19 case numbers, Time-Lagged Cross-Correlation (TLCC) between each of these signals and COVID-19 case numbers was calculated (see Correlation Analysis in Experimental). Essentially, correlations between two time series signals were repeatedly examined when one signal was incrementally shifted. If the peak correlation is at the center (offset = 0), two signals are most synchronized at the same time. However, the peak correlation may be at a different offset if one signal leads another. While there were large variations for the offsets between signals derived from news and COVID-19 cases, the offsets of GT of "COVID testing", "covid rapid testing", "COVID symptoms", and "lockdown", and COVID-19 related news volume ("news count") with COVID-19 case numbers were generally consistent and had median values of -3 to -1 across different states (Figs. 3A–G and S6), indicating a leader-follower relationship between these signals and COVID-19 cases. With small sample sizes and inconsistent samplings, wastewater COVID-19 measurement ("mean_copies_per_liter") in three states synchronized with the COVID-19 case number, consistent with the previous study [15]. These analyses indicated that GT of "COVID testing", "covid rapid testing", "COVID symptoms", and "lockdown", and news volume ("news count") could potentially serve as early predictors for COVID-19 cases.

4.4. Predicting future COVID-19 case numbers with the Prophet model

To apply these early predictors into a forecast model, single and multiple time series signals were shifted and used as extra regressor(s) in the Prophet model to predict future COVID-19 trends (see Predict COVID-19 trends with Prophet model in Experimental). Fig. 4A demonstrated the mean absolute percentage errors (MAPEs) of predicting COVID-19 cases in one week in each state using Prophet models incorporating different signals. As shown in Fig. 4B, though including some signals as single extra regressor did not improve the model's prediction accuracy, the error rate was reduced when all chosen signals were integrated into the model. When only two signals (GT of "COVID testing" and "COVID symptoms") that led to a smaller MAPE as a single extra regressor were used, the mean error rate was almost the same with that of the model using all metrics. However, in states with large mean absolute errors (MAEs) like California, Ohio, Tennessee, Illinois, Massachusetts, and Arizona, the model using two selected signals performed much worse than the model including all signals (Fig. S7). Including wastewater measurements in the model could significantly improve the model's accuracy even though the data was limited to three states (Fig. 4C). The MAPEs of predicting COVID-19 cases in different number of weeks were shown in Figs. 4D and S8. Although the mean prediction errors generally became larger as the time span of prediction increased, the median MAPE of two weeks were the same with that of one week. More importantly in many states like California, Texas, and Florida that had large number of COVID-19 cases, MAPEs of the prediction in two weeks were much smaller compared to in one week. These results demonstrated that these signals could be used to predict COVID-19 cases two weeks in advance, consistent with the results of synchrony analysis.

5. Discussion

In this article, we aimed to extract more granular signals from available public data to forecast future SARS-CoV-2 spread. We built a simple NLP pipeline that could accurately extract news of a specific topic. With the pretrained Word2Vec embeddings, this

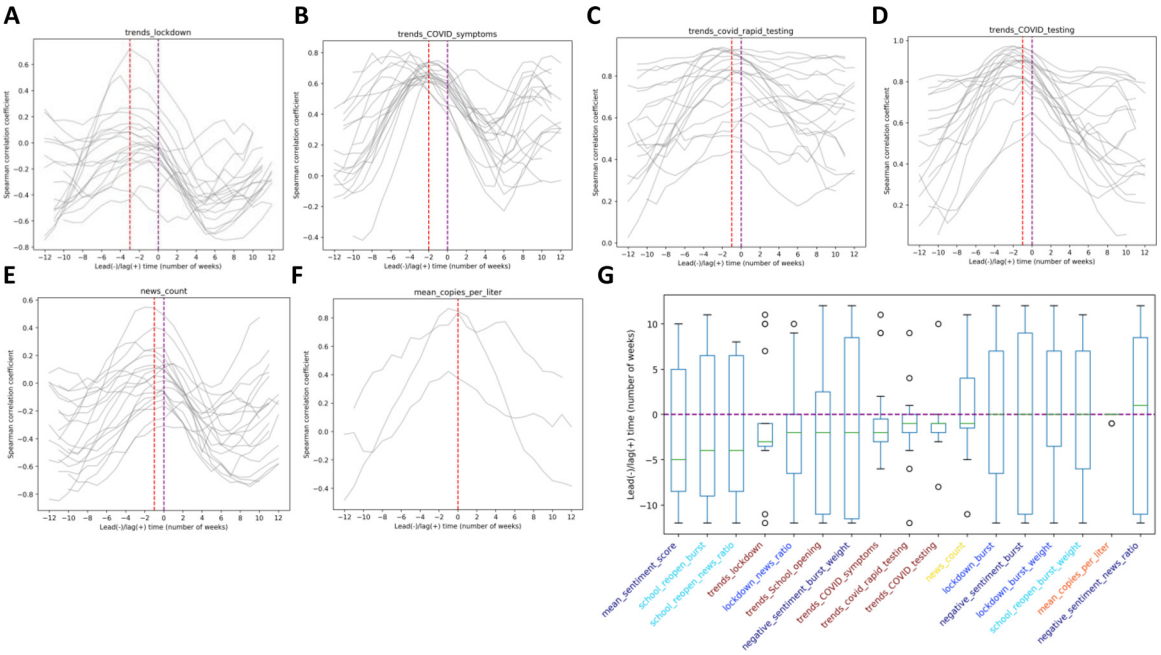


Fig. 3. Lead-lag correlation of weekly COVID-19 cases with some representative signals across different states. (A-F) Correlation coefficients with different offsets between two time-series measurements. Each line represents a state. (G) Boxplot of offsets that generate the maximum correlation coefficient across 19 states. Similar signals are labeled with the same color. Purple dashed line represents offset of 0. Red dashed line represents the median offset across different states.

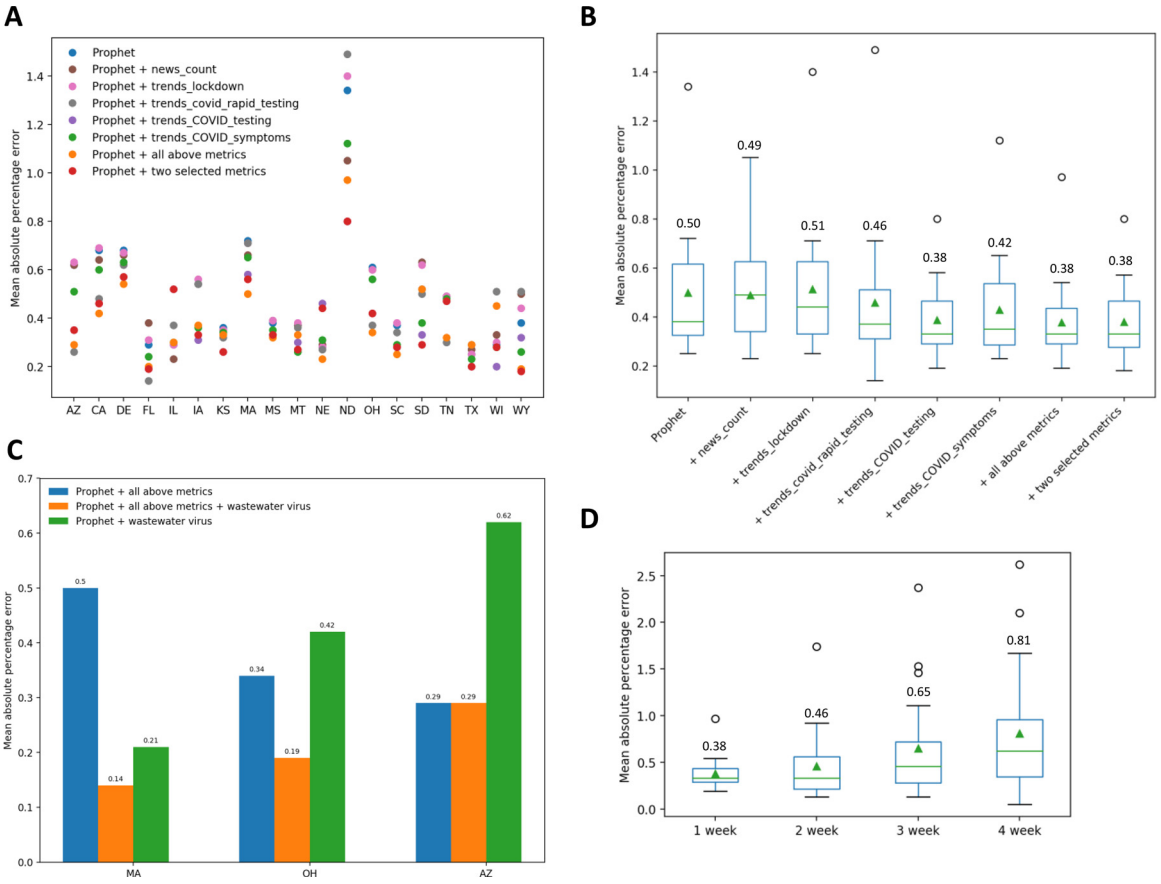


Fig. 4. Predicting future COVID-19 cases with selected metrics using the Prophet model. (A) Mean absolute percentage errors (MAPEs) of predicting COVID-19 case in one week with different Prophet models in each state. (B) Boxplot of MAPEs in (A) grouped by different Prophet models. (C) Barplot of MAPEs in predicting COVID-19 cases in one week using the Prophet model that incorporates different metrics including wastewater COVID-19 measurements. (D) Boxplot of MAPEs of predicting COVID-19 cases in 1, 2, 3, and 4 weeks across different states using the Prophet model + all above metrics. Green triangles and numbers in (B) and (D) represent the mean MAPE within each distribution.

pipeline could identify any specific news item, including public sentiment on various containment policies, and build a time series dataset for that topic, which is novel and extensible for a variety of scenarios. On top of these time series signals, we applied a burst model to mark bursty time points when special events broke out. From the signal of news with negative sentiment the burst model successfully captured the time point when the first confirmed death from COVID-19 was reported in several states (data not shown). We did not find good correlations between the signals obtained from news data with COVID-19 cases, indicating that there were many other confounding factors underlying these signals. For example, the amount of news was small even for large states due to the strict filtering step (i.e., the title must contain the state name and COVID-19 related terms) in obtaining regional COVID-19 news. In the future, we may use alternative approaches such as a classification model [27] to obtain more news data. In addition, this pipeline can be applied to other related but more abundant data sources like Twitter and Facebook posts to extract informative signals.

In contrast to news signals, Google Trends (GT) correlated well with COVID-19 case numbers and, more importantly, led 1-2 weeks ahead of the COVID-19 case numbers in almost every state that examined, consistent with previous studies [13,14]. Interestingly, the lag between GT and COVID-19 case number has a very similar range with the virus incubation period [28], suggesting people may search for COVID-19 because they suspect that they may have been exposed to the virus even though they are asymptomatic. We incorporated some GT signals as extra regressors in the Prophet model to predict the future COVID-19 case numbers, and found they significantly improved the model's performance. Based on these results, it would be worthwhile to integrate these signals with other COVID-19 forecasting models [29,30].

With limited data, we also demonstrated that the signal from wastewater COVID-19 measurement aligned with COVID-19 case number and could further reduce the Prophet model's prediction errors, especially in Massachusetts, which provided an adequate and consistent number of measurements. As more states/counties have been applying this technology [17], more data will be generated and reported, providing another valuable dataset that can be incorporated into the COVID-19 forecasting model.

As this framework is scaled up to cover more regions or larger time spans, one issue that we might confront is the sparse time series data, especially for news signals. For example, some regions might be under-reported in the news. One potential solution is to make the data less granular by deriving time sequence using month and week instead of actual date, though the resultant prediction will also be less granular and hence less useful. Another solution worth exploring is predicting missing data with latent factor models that have been commonly used in the recommender system to quantify user-item preference [31,32]. This will enable many further analyses, though the prediction of these missing values must be carefully validated.

Here, we integrated various signals into an additive Prophet forecasting model, which might not make the best use of these data. For example, bursty signals as a binary signal did not correlate well with continuous COVID-19 case numbers and thus were not included as a regressor in the model. Therefore, a new model specific for integrating these signals like [33] may generate more accurate predictions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Jose Zamalloa, Peter Z. Shen for the discussion on the methods and results. We thank Alexandra Jacunski for revising the manuscript carefully. This research is supported by Janssen Research & Development.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2022.04.030.

References

[1] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect. Dis. 20 (5) (2020) 533–534, doi:10.1016/S1473-3099(20)30120-1.
[2] G. Eysenbach, Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet, J. Med. Internet Res. 11 (1) (2009) e1157, doi:10.2196/jmir.1157.
[3] M. Farhadloo, K. Winneg, M.P.S. Chan, K.Hall Jamieson, D. Albarracin, Associations of topics of discussion on twitter with survey measures of attitudes, knowledge, and behaviors related to zika: probabilistic study in the United States, JMIR Public Health Surveill. 4 (1) (2018), doi:10.2196/publichealth.8186.
[4] L. Samaras, E. García-Barriocanal, M.-A. Sicilia, Comparing social media and google to detect and predict severe epidemics, Sci. Rep. 10 (1) (2020) 1 Art, doi:10.1038/s41598-020-61686-9.
[5] C. Poletto, P.Y. Boëlle, V. Colizza, Risk of MERS importation and onward transmission: a systematic review and analysis of cases reported to WHO, BMC Infect. Dis. 16 (1) (2016) 448, doi:10.1186/s12879-016-1787-5.
[6] G. Eysenbach, SARS and population health technology, J. Med. Internet Res. 5 (2) (2003), doi:10.2196/jmir.5.2.e14.
[7] A. Mavragani, G. Ochoa, The internet and the anti-vaccine movement: tracking the 2017 eu measles outbreak, Big Data Cogn. Comput. 2 (1) (2018) 1 Art, doi:10.3390/bdcc2010002.
[8] L.G. van Lent, H. Sungur, F.A. Kunneman, B. van de Velde, E. Das, Too far to care? Measuring public attention and fear for ebola using twitter, J. Med. Internet Res. 19 (6) (2017), doi:10.2196/jmir.7219.
[9] F. A. Binti Hamzah, et al., CoronaTracker: world-wide COVID-19 outbreak data analysis and prediction, Mar (2020), doi:10.2471/BLT.20.255695.
[10] M. Hoffman, D. Blei, F. Bach, Online Learning for Latent Dirichlet Allocation 23 (2010) 856–864.
[11] A. Mavragani, G. Ochoa, K.P. Tsagarakis, Assessing the methods, tools, and statistical approaches in google trends research: systematic review, J. Med. Internet Res. 20 (11) (2018) e9366, doi:10.2196/jmir.9366.
[12] A. Mavragani, K. Gkillas, COVID-19 predictability in the United States using Google Trends time series, Sci. Rep. 10 (1) (2020) 1 Art, doi:10.1038/s41598-020-77275-9.
[13] S.J. Kurian, et al., Correlations between COVID-19 cases and google trends data in the United States: a state-by-state analysis, Mayo Clin. Proc. 95 (11) (2020) 2370–2381, doi:10.1016/j.mayocp.2020.08.022.
[14] M. Effenberger, A. Kronbichler, J.I. Shin, G. Mayer, H. Tilg, P. Perco, Association of the COVID-19 pandemic with internet search volumes: a google TrendsTM analysis, Int. J. Infect. Dis. 95 (2020) 192–197, doi:10.1016/j.ijid.2020.04.033.
[15] J. Peccia, et al., Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics, Nat. Biotechnol. 38 (10) (2020) 10 Art, doi:10.1038/s41587-020-0684-z.
[16] X. He, et al., Temporal dynamics in viral shedding and transmissibility of COVID-19, Nat. Med. 26 (5) (2020) 5 Art, doi:10.1038/s41591-020-0869-5.
[17] F. Wu, et al., SARS-CoV-2 titers in wastewater foreshadow dynamics and clinical presentation of new COVID-19 cases, medRxiv (2020), doi:10.1101/2020.06.15.20117747.
[18] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, M. Ciccozzi, Application of the ARIMA model on the COVID-2019 epidemic dataset, Data Br. 29 (2020) 105340, doi:10.1016/j.dib.2020.105340.
[19] F.M. Khan, R. Gupta, ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India, J. Saf. Sci. Resil. 1 (1) (2020) 12–18, doi:10.1016/j.jnlssr.2020.06.007.
[20] A.K. Sahai, N. Rath, V. Sood, M.P. Singh, ARIMA modelling & forecasting of COVID-19 in top five affected countries, Diabetes Metab. Syndr. Clin. Res. Rev. 14 (5) (2020) 1419–1427, doi:10.1016/j.dsx.2020.07.042.
[21] Q. Yang, et al., Short-term forecasts and long-term mitigation evaluations for the COVID-19 epidemic in Hubei Province, China, Infect. Dis. Model. 5 (2020) 563–574, doi:10.1016/j.idm.2020.08.001.
[22] K.K. Singh, S. Kumar, P. Dixit, M.K. Bajpai, Kalman filter based short term prediction model for COVID-19 spread, Appl. Intell. 51 (5) (2021) 2714–2726, doi:10.1007/s10489-020-01948-1.
[23] S.J. Taylor, B. Letham, Forecasting at scale, PeerJ Inc. (2017) e3190v2, doi:10.7287/peerj.preprints.3190v2.

[24] C.B. Aditya Satrio, W. Darmawan, B.U. Nadia, N. Hanafiah, Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET, Procedia Comput. Sci. 179 (2021) 524–532, doi:10.1016/j.procs.2021.01.036.
[25] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *ArXiv13013781 Cs*, Sep. 2013, Accessed: Mar. 22, 2021. [Online]. Available: http://arxiv.org/abs/1301.3781.
[26] J. Kleinberg, Bursty and hierarchical structure in streams, Data Min. Knowl. Discov. 7 (4) (2003) 373–397, doi:10.1023/A:1024940629314.
[27] C. Shen, A. Chen, C. Luo, J. Zhang, B. Feng, W. Liao, 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. 22 (5) (2020) e19421, doi:10.2196/19421.
[28] S.A. Lauer, et al., The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application, Ann. Intern. Med. (2020), doi:10.7326/M20-0504.
[29] M. L. Li, H. T. Bouardi, O. S. Lami, T. A. Trikalinos, N. K. Trichakis, and D. Bertsimas, "Forecasting COVID-19 and analyzing the effect of government interventions," *medRxiv*, 2020, doi:10.1101/2020.06.23.20138693.
[30] S.O. Arik, et al., Interpretable sequence learning for COVID-19 Forecasting, ArXiv200800646 Cs Stat (Jan. 2021) Accessed: Mar23, 2021[Online]Available http://arxiv.org/abs/2008.00646 .
[31] Y. Yuan, Q. He, X. Luo, M. Shang, A multilayered-and-randomized latent factor model for high-dimensional and sparse matrices, IEEE Trans. Big Data (2020) 1–2, doi:10.1109/TBDATA.2020.2988778.
[32] M. Shang, Y. Yuan, X. Luo, M. Zhou, An α -β -divergence-generalized recommender for highly accurate predictions of missing user preferences, IEEE Trans. Cybern. (2021), doi:10.1109/TCYB.2020.3026425.
[33] D. Liu, et al., A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models, ArXiv200404019 Cs Q-Bio Stat (Apr. 2020) Accessed: Mar22, 2021[Online]Available http://arxiv.org/abs/2004.04019 .