Supporting Information for "Use Internet Search Data to Forecast COVID-19 Situation: A Systematic Review"

Simin Ma [1], Yan Sun [1] and Shihao Yang [1,*]

[1] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology
* Correspondence: shihao.yang@isye.gatech.edu

1. Sample Dataset Showcase

1.1 COVID-19 Related Data Showcase

This section showcases the COVID-19 related data discussed in section 3.1. Table S1 displays COVID-19 new case and death counts in daily frequencies for a few example countries and example dates. Here, we use the raw data from JHU CSSE COVID-19 dataset [57]. As an example, "standing on" 2022-08-07, to produce 1-4 weeks ahead forecasts (2022-08-08 to 2022-09-04), one can use COVID-19 cases and deaths information up to 2022-08-07.

Table S2 showcases the COVID-19 daily new hospital admissions (hospitalization) in United States, which is obtained from U.S. Department of Health and Human Services (HHS) [60]. Other countries' hospitalization data follows the same structure. However, the data source for difference countries varies.

Table S1. Example COVID-19 new cases and deaths dataset. The included countries are considered in the forecasting studies in Table 1. All other countries' cases and deaths data follow the same structure as shown in Table S1. The dataset source is JHU CSSE COVI-19 dataset [57].

Date	COVID-19 Data	Countries						
		China	India	Iran	Italy	United States	United Kingdom	…
2020-03-28	New Cases	56	100	3076	5974	22164	2822	…
	New Deaths	3	4	139	889	713	516	…
2020-03-29	New Cases	43	37	2901	5217	16127	2857	…
	New Deaths	5	3	123	756	555	586	…
2020-03-30	New Cases	34	227	3186	4050	22154	4273	…
	New Deaths	4	5	117	812	707	711	…
2020-03-31	New Cases	1587	146	3110	4053	26381	4515	…
	New Deaths	1	3	141	837	1079	831	…
…	…	…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…
2022-08-07	New Cases	861	0	5477	26656	12601	0	…
	New Deaths	0	0	63	74	31	98	…
…	…	…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…

Table S2. Example COVID-19 hospitalization dataset. All other countries' hospitalization data follows the same structure as shown in Table S2. But the hospitalizations' data sources vary from country to country. The dataset source is U.S. HHS [60].

Date	Countries	
	United States	…
2020-03-28	4509	…
2020-03-29	5452	…
2020-03-30	5275	…
2020-03-31	5383	…
…	…	…
2022-08-07	5400	…
…	…	…

1.2 Internet Search Data Showcase

This section showcases the internet search data discussed in section 3.2. Specifically, Google Trends' [62] search queries' frequencies are displayed below for generic internet search data illustration. Table S3 displays the search frequencies of "Loss of Taste", obtained in daily frequencies in two U.S. states. Other time frames and geographical resolutions can be specified when retrieving a query of interest. For forecasting, as an example, "standing on" 2022-08-07, to produce 1-4 weeks ahead forecasts (2022-08-08 to 2022-09-04) for COVID-19 targets, one can use the below Google Trends' information up to 2022-08-07. The dimensionality of the internet search data depends on the number of selected queries of interest.

Table S3. Example Google Trends' dataset. All other geographical resolutions' search queries' frequencies are in similar data structure. The dataset source is Google Trends [62].the ar

Date	United States	
	Georgia	California
	Loss of Taste	Loss of Taste
2020-03-28	17.69	208.95
2020-03-29	167.59	278.81
2020-03-30	234.04	127.03
2020-03-31	9.66	147.87
…	…	…
2022-08-07	54.18	106.53

…	…	…

2. More Details on Forecsating Methods

2.1 More Details on Baseline LSTM Architecture

This section further demonstrates the baseline LSTM architecture that is incorporated by Ayyoubzadeh et al. [46] and Prasanth et al. [47]. The vanilla LSTM architecture is implemented using four gates: cell, input, forget and output gates. The working equations of the LSTM gates are shown as follows [77]:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \tag{1}$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \tag{2}$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \tag{3}$$

$$\tilde{c}_t = \sigma(W_c[h_{t-1}, x_t] + b_c), \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_{t-1}, \tag{5}$$

$$h_t = o_t \odot \tanh(c_t), \tag{6}$$

where f_t is the output of forget gate for time t, h_{t-1} is the output of the hidden state vector form time $t-1$, i_t represents the input state at time t, o_t represents the output state at time t, c_t represents the output from the cell unit, W_i, W_c, W_f and W_o are the weight matrices associated with the input, cell unit, forget and output gates, while b_i, b_c, b_f and b_o are the bias vectors associated with input, cell unit, forget and output gates, σ represents the sigma activation function and $\odot$ represents the Hadamard operation on two matrices. The architecture of the LSTM cell with all the gates and variables are also shown in Figure S1.

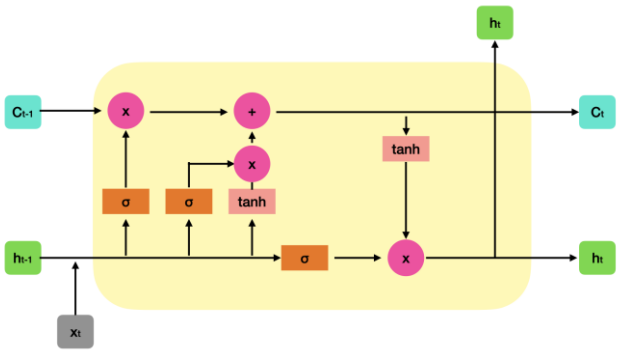

Figure S1. Illustration of the architecture of a baseline LSTM cell.

2.2 More Details on Statistical Models

This section further expands on the traditional statistical model structures for time series forecasting. In disease forecasting task, one of the most used linear based models is auto-regressive moving average model (ARMA), and its variants, due to its simplicity and ex-

plainability. In this section, we briefly expand on an ARMA-like structure, AutoRegressive and GOogle search data (ARGO) [12], and its variant proposed for COVID-19 forecast by Ma and Yang [51]. Other variants of ARGO is proposed by Liu et al. [45], Wang et al. [52], and Ma et al. [53]. Consider the following forecasting scenario:

Let $X_{j,t}$ be the Google Trends data of search term j on day t; y_t be the COVID-19 target at day t. Inspired by ARGO method [12], Ma and Yang [51] proposed to combine the lagged COVID-19 target and internet search data for future COVID-19 target's predictions. Specifically, with information available as of time T, to estimate y_{T+l} for $l > 0$, the COVID-19 target on day $(T + l)$, an L_1 regularized linear estimator is used:

$$\hat{y}_{T+l} = \hat{\mu}_y + \sum_{i=0}^{I} \hat{\alpha}_i \, y_{T-i} + \sum_{j=1}^{J} \hat{\beta}_j \, X_{j,t}$$

where one can use lagged COVID-19 target and latest Google search information. For l^{th} day ahead prediction, the coefficients $\{\mu_y, (\alpha_0, \dots, \alpha_I), (\beta_1, \dots, \beta_J)\}$ can be obtained via

$$\underset{\mu_y, \alpha, \beta, \lambda}{\operatorname{argmax}} \sum_{t=T-N-l+1}^{T-l} \left(y_{t+l} - \mu_y - \sum_{i=0}^{I} \alpha_i \, y_{t-i} - \sum_{j=1}^{J} \beta_j \, X_{j,t} \right)^2 + \lambda_\alpha ||\boldsymbol{\alpha}||_1 + \lambda_\beta ||\boldsymbol{\beta}||_1$$

where N is the length of training period; I is the total lags of the COVID-19 target considered; $\mathcal{J}$ is the total number of relevant queries considered. One can set hyperparameters $\lambda = (\lambda_\alpha, \lambda_\beta, \lambda_\delta, \lambda_\gamma)$ through cross-validation. For simplicity, Ma and Yang [51] suggest to constrain $\lambda_\alpha = \lambda_\beta = \lambda_\delta = \lambda_\gamma$. Ma and Yang [51] has detailly demonstrated that the L_1 regularization term plays a crucial role in COVID-19 forecast, as it can robustly select the important queries among a large poll for accurate COVID-19 deaths' predictions in different geographical resolution. One can also include other exogenous variables and consider different lagging periods for various rapidly changing COVID-19 dynamics.

3. Accessibility of Selected Research Studies

Table S4 below detailly collects the data and code availabilities of all the selected research studies considered in this review paper. For those studies that do not have data or code availability stated in the manuscript, "N/A" is used. Table S4 provides a systematic look-up table to provide accessibility and reproducibility for the selected studies.

Table S4. Data and Code availabilities of the selected research studies in this review paper.

Study	Data Availability	Code Availability
Liu et al. (2020) [45]	Harvard Dataverse, DOI: https://doi.org/10.7910/DVN/MWMLDV	Harvard Datavserse, DOI: https://doi.org/10.7910/DVN/MWMLDV
Ayyoubzadeh et al. (2020) [46]	N/A	N/A
Prasanth et al. (2021) [47]	N/A	N/A
Rabiolo et al (2021) [48]	Rabiolo A, Alladio E, Morales E, Marchese A. PredictPandemic.org. URL: https://predictpandemic.org [accessed 2022-11-12]	Rabiolo A, Alladio E, Morales E, Marchese A. PredictPandemic web-application. URL: https://zenodo.org/record/4603713#.YE4sfC2l28U [accessed 2022-11-12]
Lampos (2021) [49]	https://figshare.com/projects/Tracking_COVID-19_using_online_search/81548	N/A

Turk et al. (2021) [50]	Github Repository: https://github.com/philturk/CovCenVECM.git	Github Repository: https://github.com/philturk/CovCenVECM.git
Ma and Yang (2022) [51]	Harvard Dataverse, DOI: https://doi.org/10.7910/DVN/IBJS6X	Github Repository: https://github.com/stevenmsm/COVID-19-Forecasts-Using-Internet-Search-Information-in-the-United-States.git
Wang et al. (2022) [52]	Harvard Dataverse, DOI: https://doi.org/10.7910/DVN/S7HOTD	Harvard Dataverse, DOI: https://doi.org/10.7910/DVN/S7HOTD
Ma et al (2022) [53]	Harvard Dataverse, DOI: https://doi.org/10.7910/DVN/PGNBAX	Github Repository: https://github.com/stevenmsm/Joint-COVID-19-and-Influenza-Forecasts-in-the-United-States.git