Sentiment Informed Timeseries Analyzing AI (SITALA) to curb the spread of COVID-19 in Houston

Author: Prathamesh S. Desai^{1*}

Affiliations:

¹Mechanical Engineering, W. M. Rice University, Houston, TX.

*Correspondence to: pdesai@rice.edu (PSD); Website: <https://sites.google.com/site/pratnsaiweb/>

Abstract: Coronavirus disease (COVID-19) has evolved into a pandemic with many unknowns. Houston, located in the Harris County of Texas, is becoming the next hotspot of this pandemic. With a severe decline in international and inter-state travel, a model at the county level, as opposed to the state or country level, is needed. Existing approaches have a few drawbacks. Firstly, the data used is the number of COVID-19 positive cases instead of positivity. The former is a function of the number of tests carried out while the latter is normalized by the number of tests. Positivity gives a better picture of the spread of this pandemic as with time more tests are being administered. Positivity under 5% has been desired for the reopening of businesses to almost 100% capacity. Secondly, the data used by models like SEIRD lacks information about the sentiment of people with respect to coronavirus. Thirdly, models that make use of social media posts might have too much noise. News sentiment, on the other hand, can capture long term effects of hidden variables like public policy, opinions of local doctors, and disobedience of state-wide mandates. The present study introduces a new AI model, viz., Sentiment Informed Timeseries Analyzing AI (SITALA), that has been trained on COVID-19 test positivity data and news sentiment from over 2750 news articles for the Harris county. The news sentiment was obtained using IBM Watson Discovery News. SITALA is inspired by Google-Wavenet architecture and makes use of TensorFlow. The mean absolute error for the training dataset of 66 consecutive days is 2.76 and that for the test dataset of 22 consecutive days is 9.6. The model forecasts that in order to curb the spread of coronavirus in Houston, a sustained negative news sentiment will be desirable. Public policymakers may use SITALA to set the tone of the local policies and mandates.

One Sentence Summary: An AI model based on news sentiment and COVID-19 test positivity is developed to predict the spread of coronavirus in Houston and to guide future local public policy.

Background, problem, and solution In the present-day USA, pandemic models of COVID-19 at the level of state or country (1-4) are not of much use due to a severe decline in air travel (5,6). The most predominant spread of the virus is then restricted to the geography of a county or a couple of neighboring counties. National, and more so local, news articles provide a fairly accurate picture of the ongoing situation in a crisis-stricken county, Harris in the case of the present study. Local public policy makers can thus play an important role to set the tone or sentiment of the news at the county level (7). The spread of coronavirus in the short-term future is a function of test positivity data, however, over longer time periods, news sentiment starts playing an important role. For example, a stay-at-home order, with a strong negative sentiment, issued today may only start seeing the decline in test positivity after 10-14 days due to the incubation period of the coronavirus. Existing approaches have a few drawbacks (2,8,9,10): Firstly, the data used is the number of COVID-19 positive cases instead of positivity. Secondly, the data used by models like SEIRD

lacks information about the sentiment of people with respect to coronavirus. Thirdly, models which make use of social media posts might have too much noise. This study attempts to develop a multivariate artificial intelligence (AI) model to analyze timeseries of COVID-19 positivity and news sentiment. The AI model is inspired by Google's Wavenet (11) architecture and uses IBM Watson Discovery News (12) to mine COVID-19 sentiment in the news articles.

Data: The COVID-19 test positivity data for Harris county was obtained from the website of Texas Department of State Health Services (<https://dshs.texas.gov/coronavirus/additionaldata.aspx>). Couple of instances of bad or missing data were filled using linear interpolation. IBM Watson Discovery was used to mine the news sentiment in **2867 news articles** over the period of 3 months. The query used in this study is provided in the supplementary material. The entire dataset is also provided in the supplementary material.

Model (11-16): Daily COVID-19 positivity rate and news sentiment are passed to a Wavenet-inspired multivariate CNN to predict the future COVID-19 positivity rate. The AI is named **SITALA** or Sentiment Informed Timeseries Analyzing AI. A 16-day window, based on coronavirus incubation period of 14 days, with a stride of 1 was used to generate training and test datasets. The architecture of SITALA is shown in **Fig. 1**. Dilated causal convolutions help with the transmission of long-term effects. The output is a single point prediction of COVID-19 in the future. Python code for the neural network architecture is provided in the supplementary material.

Fig. 1. Architecture of SITALA: It is a sequential model that takes a window of multivariate (viz., COVID test positivity and news sentiment from IBM Watson News Discovery) timeseries and outputs the COVID test positivity at the next timestep. This architecture is inspired by Google's Wavenet (11) architecture (viz., dilated causal convolutions). In the present study, a window size of 16 was used and dilations of 1, 2, 4, 8 were used.

Results: The data used in the present study is shown in **Fig. 2** along with the predictions of SITALA. Data for the Harris county from 04/21 to 07/17 has been used in this study. Number of news articles returned by IBM Watson Discovery News is shown with bars that are multiplied by the sign of the average daily news sentiment. The average daily news sentiment (connected blue

squares) can be +1.0 for maximum positive sentiment and -1.0 for maximum negative sentiment. Overall, on most of the days the news sentiment about the spread of coronavirus in Houston, Harris county, Texas has been negative. A big positive spike in news sentiment is seen around the time of social unrest in Houston (05/30 to 06/02). The focus of news might have shifted away from COVID-19 during this timeframe. An upward trend is visible in the COVID-19 positivity data (connected red dots).

Fig. 2. SITALA predictions and relevant COVID-19 data for the Harris county, Texas. Number of news articles returned by IBM Watson discovery, shown on the right axis, started increasing from around mid-May, 2020. COVID-19 test positivity and news sentiment are shown on the left axis. Around 75% of the data (green window) was used for training SITALA, of which 10% was reserved for validation. SITALA was tested on remaining 25% of the data (blue window) for which the mean absolute error (MAE) was 9.6. SITALA forecast (gray window) shows how maintaining a negative sentiment in the news about the spread of COVID-19 can be beneficial to control and eventually decrease test positivity.

The data was divided into training and test datasets comprised of 75% and 25% of the entire data (i.e., 04/21 to 07/17) respectively. 10% of the training data was used for validation. The continuous black line with shadow shows the predictions of trained SITALA over the entire dataset. SITALA is able to capture the response of the COVID-19 positivity data with a mean absolute error (MAE) of 2.76 for training dataset and 9.6 for test dataset. SITALA is unable to capture the highest spikes encountered in both the datasets of COVID-19 positivity. This may have been due to the smaller number of observations in the total dataset (a total of 88 days' worth of observations is not at the level of the requirements of big data) and smoothing out effect that is introduced by the time window of 16 days.

A maximum positive sentiment of 0.7 and negative of -0.7 were used as the bounds for future news sentiment. With these as the inputs to SITALA, two extreme forecasts were obtained till 08/07. These are also shown in **Fig. 2** with dotted red and dotted green lines with shadow, respectively. SITALA forecasts the positivity of COVID-19 in Houston to lie within this uncertainty cone. A sustained positive sentiment, e.g., “masks are optional”, may prove disastrous for the spread of coronavirus in Houston. On the contrary, a sustained negative sentiment, e.g., “death count for COVID-19 is growing at an alarming rate in Houston”, may help to discourage social gatherings and to keep the COVID-19 positivity under check.

Discussion: This study highlights the multivariate nature of COVID-19 positivity. The unknowns about the disease have not yet been thoroughly understood. However, public policy makers can benefit from models like SITALA which add the dimension of news sentiment to the positivity data to make forecasts. The long-term effect of sentiment due to the virus incubation period of 14 days can be captured using an AI making use of dilated causal convolutions. News publishers also have a bigger role to play in curbing the spread of coronavirus in Houston. SITALA is a constantly evolving AI and should be enhanced with newer data, as and when available, using transfer learning. SITALA may be deployed at other similar crisis-stricken counties in New York, Florida, and California.

Limitations: The query searched for the articles having ‘houston’ in the url may have caused omission of few relevant articles that did not have Houston in the url. During the initial few days of training dataset, there were hardly any articles relevant to the IBM Watson query and thus the sentiment during this period was assumed to be neutral, i.e., a value of 0.

Ethics (17, 18): SITALA may only be used as an ethical guide by public policy makers to set the tone of policies. The author does not support any unethical use of SITALA.

References and Notes:

1. Giuliani, Diego, et al. "Modelling and predicting the spatio-temporal spread of Coronavirus disease 2019 (COVID-19) in Italy." *Available at SSRN 3559569* (2020).
2. Yang, Zifeng, et al. "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions." *Journal of Thoracic Disease* 12.3 (2020): 165.
3. Benvenuto, Domenico, et al. "Application of the ARIMA model on the COVID-2019 epidemic dataset." *Data in brief* (2020): 105340.
4. Dandekar, Raj, and George Barbastathis. "Neural Network aided quarantine control model estimation of global Covid-19 spread." *arXiv preprint arXiv:2004.02752* (2020).
5. United State Department of Transportation "Air Traffic Data, May 2020: 89% Reduction in U.S. Airline Passengers from May 2019 (Preliminary)" (Available online: <https://www.bts.gov/newsroom/air-traffic-data-may-2020-89-reduction-us-airline-passengers-may-2019-preliminary>) (2020) [the easiest access to this source is via the URL]
6. Chinazzi, Matteo, et al. "The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak." *Science* 368.6489 (2020): 395-400.
7. Adolph, Christopher, et al. "Pandemic politics: Timing state-level social distancing responses to COVID-19." *medRxiv* (2020).
8. Shinde, Gitanjali R., et al. "Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art." *SN Computer Science* 1.4 (2020): 1-15.
9. Karisani, Negin, and Payam Karisani. "Mining Coronavirus (COVID-19) Posts in Social Media." *arXiv preprint arXiv:2004.06778* (2020).
10. Ayyoubzadeh, Seyed Mohammad, et al. "Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study." *JMIR Public Health and Surveillance* 6.2 (2020): e18828.
11. Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).
12. High, Rob. "The era of cognitive systems: An inside look at IBM Watson and how it works." *IBM Corporation, Redbooks* (2012): 1-16.
13. Géron, Aurélien. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.
14. Ting, Daniel Shu Wei, et al. "Digital technology and COVID-19." *Nature medicine* 26.4 (2020): 459-461.
15. Wan, Renzhuo, et al. "Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting." *Electronics* 8.8 (2019): 876.
16. Borovykh, Anastasia, Sander Bohte, and Cornelis W. Oosterlee. "Conditional time series forecasting with convolutional neural networks." *arXiv preprint arXiv:1703.04691* (2017).
17. Cutler, Adam, Milena Pribić, and Lawrence Humphrey. "Everyday ethics for artificial intelligence." *PDF, IBM Corporation* (2019).
18. Ienca, Marcello, and Effy Vayena. "On the responsible use of digital data to tackle the COVID-19 pandemic." *Nature medicine* 26.4 (2020): 463-464.

Funding: No specific funding was received for this work; **Author contributions:** PSD designed the research, performed the research, analyzed the data, wrote and edited the manuscript; **Competing interests:** PSD declares no competing interests; and **Data and materials availability:** All data is available in the main text or in the supplementary materials.

Supplementary Materials:

Material and Methods

Fig. S1

Table S1

Materials and methods

COVID-19 test positivity data This data is available to the general public and can be downloaded from the website of Texas Department of State Health Service (<https://dshs.texas.gov/coronavirus/additionaldata.aspx>).

News sentiment data The data for news sentiment was obtained by querying IBM's Watson Discovery News tool. The tool provides 200 free queries per user per month. More information can be found here: <https://www.ibm.com/watson/services/discovery-news/>. Discovery News employs natural language processing to return answers to the queries. It also analyzes the sentiment of the news articles. The query used in this study was:

Is the spread of coronavirus or covid-19 or 2019-nCoV under control?

Include analysis of your results

`average(enriched_text.sentiment.document.score)`

Filter which documents you query

`publication_date::"2020-05-29",url:"houston", (enriched_text.keywords.text:"coronavirus"|enriched_text.keywords.text:"COVID-19"|enriched_text.keywords.text:"2019-nCoV")`

A sample query along with the output from Watson Discovery News is shown in **Fig. S1**.

Fig. S1. Sample query output from IBM Watson News Discovery. The present study used the exact same query to determine the sentiment of news (-1 implies maximum negative sentiment and +1 implies maximum positive sentiment) over varying publication dates. Location-specific (viz., Houston) articles were filtered using— `url:"houston"`

The entire dataset is reproduced in **Table S1**

Table. S1. Entire dataset, for the Harris county, that was used in the present study. Also included are the predictions and forecast of SITALA. Window size of 16 was used in this study therefore the predictions of SITALA start from 17th day (i.e., 5/7)

	Date	Inputs		Output	SITALA predictions		
		Percent cleaned test positivity	Percent COVID news sentiment	Percent cleaned test positivity	Positivity predictions	Forecast with positive sentiment (70)	Forecast with negative sentiment (-70)
Training dataset with 10% reserved for validation	4/22	10.02	0.00	10.02			
	4/23	2.24	0.00	2.24			
	4/24	3.56	0.00	3.56			
	4/25	4.38	0.00	4.38			
	4/26	21.19	0.00	21.19			
	4/27	3.93	-61.81	3.93			
	4/28	14.83	0.00	14.83			
	4/29	4.24	0.00	4.24			
	4/30	7.58	0.00	7.58			
	5/1	6.67	0.00	6.67			
	5/2	5.76	0.00	5.76			
	5/3	4.89	0.00	4.89			
	5/4	4.29	-52.66	4.29			
	5/5	3.73	0.00	3.73			
	5/6	3.18	0.00	3.18			
	5/7	2.62	0.00	2.62	2.60		
	5/8	2.36	0.00	2.36	2.49		
	5/9	6.29	0.00	6.29	6.08		
	5/10	4.20	0.00	4.20	4.16		
	5/11	2.23	0.00	2.23	2.06		
	5/12	3.93	0.00	3.93	3.67		
	5/13	4.50	0.00	4.50	4.32		
	5/14	4.29	0.00	4.29	4.17		
	5/15	9.17	5.43	9.17	8.85		
	5/16	4.78	1.36	4.78	4.59		
	5/17	1.81	-26.88	1.81	1.76		
	5/18	7.66	13.96	7.66	7.32		
	5/19	2.38	-10.04	2.38	2.19		
	5/20	5.00	7.82	5.00	4.88		
	5/21	6.64	1.89	6.64	6.49		
	5/22	8.48	11.79	8.48	8.32		
	5/23	6.74	-54.95	6.74	6.56		
	5/24	4.99	-11.66	4.99	4.92		
	5/25	3.25	-30.84	3.25	3.19		
	5/26	1.46	6.23	1.46	1.52		
	5/27	14.83	11.26	14.83	14.78		
	5/28	4.27	11.29	4.27	4.35		
	5/29	6.20	5.52	6.20	5.92		
	5/30	6.06	39.18	6.06	6.00		
	5/31	21.91	25.26	21.91	21.21		
	6/1	0.96	16.38	0.96	0.68		
	6/2	8.34	0.33	8.34	7.94		
	6/3	12.92	-3.21	12.92	12.45		
	6/4	17.55	15.25	17.55	17.02		
	6/5	12.49	9.44	12.49	11.88		
	6/6	7.89	-26.27	7.89	6.94		
	6/7	5.57	-15.18	5.57	5.21		
	6/8	11.63	-19.69	11.63	11.01		
	6/9	16.62	-14.70	16.62	16.26		
	6/10	15.01	2.63	15.01	14.11		
	6/11	13.40	-12.09	13.40	13.22		
	6/12	11.79	-18.90	11.79	11.28		
	6/13	10.19	-28.92	10.19	9.99		
	6/14	7.46	2.87	7.46	7.58		
	6/15	7.16	-4.26	7.16	7.27		
	6/16	5.53	2.51	5.53	5.57		
	6/17	10.04	14.69	10.04	10.10		
	6/18	7.08	-0.39	7.08	6.87		
	6/19	5.75	15.41	5.75	5.68		
	6/20	17.55	6.68	17.55	16.89		
	6/21	27.71	-27.42	27.71	6.39		
	6/22	3.71	4.48	3.71	6.63		
	6/23	60.35	-17.23	60.35	12.43		
	6/24	45.44	-10.17	45.44	9.54		
	6/25	30.53	-0.26	30.53	13.65		
	6/26	15.62	-16.69	15.62	11.36		
	6/27	9.05	-10.33	9.05	14.38		
	6/28	9.18	-48.70	9.18	14.08		
	6/29	1.64	-20.83	1.64	17.87		
	6/30	19.83	-15.70	19.83	11.95		
Test dataset	7/1	14.38	-20.46	14.38	9.54		
	7/2	20.69	-30.37	20.69	17.72		
	7/3	23.57	-11.68	23.57	8.66		
	7/4	15.27	-24.72	15.27	12.23		
	7/5	6.32	-35.68	6.32	16.23		
	7/6	7.38	1.79	7.38	11.14		
	7/7	13.42	3.53	13.42	12.01		
	7/8	15.12	-41.36	15.12	11.32		
	7/9	13.59	-50.11	13.59	10.55		
	7/10	11.44	-27.35	11.44	12.04		
Forecast	7/11	10.81	-21.80	10.81	5.43		
	7/12	25.92	-51.09	25.92	11.06		
	7/13	18.76	-4.39	18.76	6.50		
	7/14	49.02	0.49	49.02	6.91		
	7/15	28.24	-11.31	28.24	7.88		
	7/16	27.44	1.59	27.44	16.13		
	7/17	26.65	-9.15	26.65	9.96		
	7/18	20.28	-31.79	20.28	20.28	20.28	20.28
	7/19					9.23	9.23
	7/20					17.31	13.78
	7/21					26.54	17.34
	7/22					14.89	8.86
	7/23					22.31	11.32
	7/24					32.09	16.55
	7/25					26.15	8.79
	7/26					29.99	10.60
	7/27					37.56	14.11
	7/28					33.28	12.61
	7/29					41.10	9.85
	7/30					44.90	15.87
	7/31					45.41	14.24
	8/1					49.85	16.57
	8/2					57.47	18.08
	8/3					56.75	19.36
	8/4					55.62	16.81
	8/5					56.75	17.65
	8/6					59.65	17.22
	8/7					60.82	17.17

SITALA architecture The python code for the architecture of SITALA is reproduced below. Causal padding and relu activation were used along with 4 dilated convolutional hidden layers. Contact PSD (pdesai@rice.edu) to obtain the trained SITALA for transfer learning.

```

5  # SITALA - Sentiment Informed Timeseries Analyzing AI
   # Location - Harris county, TX
   # Purpose - Forecast spread of coronavirus
   # © 2020. Prathamesh S. Desai. All rights reserved.

10  tf.keras.backend.clear_session()
   tf.random.set_seed(40)
   np.random.seed(40)

   # Model definition
15  model = tf.keras.Sequential()
   model.add(tf.keras.layers.InputLayer(input_shape=(window, n_features)))

   for dilation_rate in (1, 2, 4, 8):
20       model.add(
           tf.keras.layers.Conv1D(filters=64,
                                   kernel_size=2,
                                   strides=1,
                                   dilation_rate=dilation_rate,
                                   padding="causal",
                                   activation="relu")
25       )

   model.add(tf.keras.layers.Flatten())
   model.add(tf.keras.layers.Dense(128, activation='relu'))
30  model.add(tf.keras.layers.Dense(1))

   # Model compilation
   optimizer = tf.keras.optimizers.Adam(lr=5e-4)
35  model.compile(loss=tf.keras.losses.Huber(),
                 optimizer=optimizer,
                 metrics=["mae"])

   early_stopping = tf.keras.callbacks.EarlyStopping(patience=200)
40  model.summary()

   history = model.fit(X_train, Y_train, epochs=500, verbose=True,
                       validation_split=0.1,
                       callbacks=[early_stopping])
45

   print("SITALA has been trained")

```