Study of inter-regional relations for various disease symptoms generated by Google Trends

Anshuman Sinha
Georgia Institute of Technology
Atlanta, US

Arvind Bangaru
Georgia Institute of Technology
Atlanta, US

Bhavay Aggarwal
Georgia Institute of Technology
Atlanta, US

ABSTRACT

The use of hospital resources and the development of management plans to best manage infected patients are dependent on accurate forecasting of the number of COVID-19 cases. Although monitoring of sensors is a good method to measure the spread of the disease, it an expensive task and has several privacy and ethical issues. A low-cost alternative to sensors in monitoring the sensors can be leveraging google search trends on disease symptoms. It is not far fetching to think that people having such symptoms would google them in order to get more information and potential remedies or cures. If this assumption is experimentally validated, it becomes possible to forecast covid-19 cases by forecasting symptoms related to covid-19. In this proposal, we look at how the disease forecasting problem has been previously approached and how google trends data can be incorporated into it.

KEYWORDS

Covid-19, Google search trends, Epidemiology, Deep learning, Time-series, Disease forecasting

ACM Reference Format:
Anshuman Sinha, Arvind Bangaru, and Bhavay Aggarwal. 2022. Study of inter-regional relations for various disease symptoms generated by Google Trends. In *Proceedings of Gatech (CSE 8803).* ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

Covid-19 is a viral contagious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus. It has spread across the globe and was declared a global pandemic in march 2020. The initial strains have evolved into more potent strains like delta and omicron which are more infectious and show more severe symptoms. Covid-19 is a flu-like disease and has similar methods of transmission. Symptoms of covid-19 Include fever or chills, cough, shortness of breath or difficulty breathing, fatigue, muscle or body aches, headache, new loss of taste or smell, sore throat, congestion, nausea, and diarrhoea among others. It is important to note that a third of the infected do not show any symptoms but can asymptomatically transmit the disease to others. Most of the infected show only mild symptoms with only a small portion developing severe pneumonia. Covid-19 is tested for by detecting the presence of SARS-CoV-2 specific DNA/antigens/antibodies. Positive test results can occur in infected persons from about five days after the initial infection to several weeks. The period in which the infected person can transmit the disease to others is far lower (only about 10 days after infection). Managing the pandemic involves collecting data in forms of surveys, Covid-19 test results and electronic health records. There are challenges to collecting and using these data like large-scale surveying being expensive, time-consuming, and having limitations in terms of capturing time-sensitive data and accuracy. Electronic health data requires complex anonymization, merging and analysis before they can be used. Due to these limitations, alternative approaches like information-seeking behaviours of the population captured in data like Google search trends are an exciting avenue for forecasting and managing pandemics. Covid-19 has had far-reaching economic, cultural, and social impacts and has led to the COVID-19 recession. Predicting and planning for local covid-19 outbreaks is very important from a public health perspective and this study helps in this prediction from publicly available data

CSE 8803, Course Project,
2022.

2 LITERATURE REVIEW

Statistical significance of correlations between Google Trends and COVID-19 data: the estimated models exhibit strong COVID-19 predictability. Several studies have been carried out using correlation and regressing models to model and forecast Covid-19 cases using Google Trends. Mavragani et al. [6] studied the association between Google Trends data and COVID-19 data on cases and fatalities and built a quantile regression model for forecasting. Similarly, Kim et al. correlated new cases of Covid-19 with Google Trends of search terms related to loss of smell and taste using loss of hearing and sight as controls. They found that only using search trends of symptoms (loss of smell and taste) is not enough for accurately predicting the levels of new cases as these may be influenced by extraneous influences like media coverage of those specific symptoms [2]. These studies show that Google Trends and COVID-19 data display statistically significant relationships, the calculated models show good COVID-19 predictability and can be used for predicting covid-19 outbreaks after they have been corrected for some extraneous influences.

Husnayain et al. [4] explored the potential correlation between particular google search trends (related to covid, sanitation and face masks) and the number of COVID-19 infections in Taiwan (from the Taiwan CDC website). The authors noted that there was a significant increase in the search trends for each of the selected terms that peaked before the peak of the covid cases and then continuously declined afterwards due to the increase in the mainstream availability of covid related health information. They also observed an increase in google searches one to 3 days prior to the increase in covid cases. Mangono et al. used Google Trends as a proxy for what people are thinking, needing, and planning in real-time across the United States in response to COVID-19 using search trends for terms falling into six themes: social and travel, care seeking, government programs, health programs, news and influence, outlook and

concerns. and found several correlations[5]. They explored how search terms varied across countries and how searching for covid-19-related terms influenced other searches related to health. These findings indicate that google search trends can be used not only to forecast the rise in covid cases but also to understand population behaviour and allow public health bodies to make more informed and tailored responses.

Yang et.al. [9] in 2010 studied the effect of environmental factors on incidence of depression using seasonality of Internet search trend of depression, geographic location, temperature and solar radiation with the help of cross correlation coefficient between the intrinsic mode functions of local search trends. They assessed the association of amplitudes between search trend IMFs and temperature in order to explore the influence of latitude on the magnitude of seasonality of search trends. Using an adaptive approach, such empirical mode decomposition (EMD) analysis [3], might be helpful in separating significant seasonal components because fluctuation in search trend data typically consists of numerous periodic components with non-stationary and nonlinear properties[8]. Similarly, we can explore the correlation between various diseases with COVID 19.

Forecasting Covid-19 cases using machine learning
The use of hospital resources and the development of management plans to best manage infected patients are dependent on accurate forecasting of the number of COVID-19 cases. Past couple of years has seen a great rise in the number of applications for machine learning and deep learning in this field! Rustom et. al. [7] explored four Machine learning model for prediction of Covid-19 cases for various countries. Zeroual et al. [10] provide a comprehensive comparison of deep learning models in order to forcast the time-series data of covid-19. This paper presents a comparative study of five deep learning methods to forecast the number of new cases and recovered cases. In order to predict the number of covid cases (new) and recovered cases, this paper compares five deep learning techniques. Based on the amount of data from six different countries, the paper compares the following NN models; Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM) , Gated Recurrent Units (GRUs), Variational AutoEncoder (VAE)

The enhanced capabilities of deep learning models to capture process non-linearity and their adaptability in modeling time-dependent data are major driving forces behind this decision. Results show that VAE performed the best and was closely followed by the LSTMs. LSTM is a sophisticated gated memory unit created to address the vanishing gradient issues that restrict the effectiveness of a straightforward RNN.

Adhikari et al. [1] proposed a deep learning approach for forecasting of influenza cases by learning latent space representations and learning from similarities with previous disease seasons. The objective of the model was divided into 4 primary tasks -

- **Future incidence** – Short term forcasting by predicting t – 1 to t + 2 wILI(weighted Influenza-like Illness) values given time series data till week t – 2.
- **Seasonal Peak Intensity** – predict the maximum intensity of influenza in the given season.
- **Seasonal Peak Week** – Predict the week of maximum intensity.
- **Onset Week** – Predict the week when the flu season starts to rise/spread.

The model consists of an encoder-decoder and a clustering module. The clustering module uses Improved Deep Embedded clustering(IDEC) which minimizes the KL divergence between the current season embedding and the historical season clusters to learn the similarity between the learnt embeddings and similar influenza seasons in history. Since our current approach does not include using previous seasons/diseases, we want to focus more on understanding the encoder-decoder module. The encoder consists of an LSTM network which, when given time series data, utilizes the sequence of incidence of cases to produce meaningful predictions. The LSTM network, however, tends to over rely on the most frequent observations and to overcome this the attention mechanism was used which learns variable weights for all previous hidden states to make more informed predictions. The output of the encoder and the clustering module is then used to make the 4 types of predictions.

3 PROBLEM FORMULATION

The overall study of inter-regional relation of google trends of various disease symptoms and Covid-19 cases is divided into the following sub-parts:

- **Correlations:** Study of co-relation between various disease symptoms (from google trends) and covid-19 cases.
- **Predictions:** Covid-19 prediction in the following cases.
  a Prediction of covid-19 cases with the help of trends of a single symptom.
  b Based on results of (a) . Prediction of covid 19 cases using trends of multiple symptoms.
- **Inference:** based on (a) and (b) predict the covid 19 cases in a state on the basis of trends of various related symptoms in the neighbouring states
- **Model improvement:** Study the impact of media coverage on google trends ability to over-predict the related diseases and eventually its effects on estimates of covid-19.

Given $\mathcal{S} = \{S_i\}_{i=1}^{N}$ be the set of all symptoms in the COVID-19 Search Trends symptoms dataset. S_i^t represents the search volume of symptom i at time t. Let $\mathcal{Y}$ be the time-series data of covid-19 infections and $\mathcal{Y}^t$ denote the reported number of covid-19 cases at time t. Our initial task is to determine whether there is any correlation between the google search trends for a particular symptom and the covid-19 cases in a given region from time $t = 0$ till time t.

$$\rho_{\mathcal{Y},S_i} = corr(\mathcal{Y}, S_i) \quad \forall\, i \in N$$

We are yet to determine how many symptoms would be important in predicting covid-19 cases but for now let's assume we pick the top k most correlated symptoms. Next, we try to forecast covid-19 cases

using each symptom individually and by using their combinations. This task will be treated as a short term forecasting task where the covid-19 cases time series data $\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2..\mathcal{Y}_{t-1}\}$ and google search trend data $S_i = \{S_i^1, S_i^2..S_i^{t-1}\} \forall\, i \in k$ till time t is input into a neural network which predicts covid-19 cases y_{t+1} upto y_{t+w} where w is the number of weeks we want to forecast. Another important task is to predict whether neighbouring states influence the trend of covid-19 cases for any given state. The first step here again is to check whether the google search trends of the symptoms are correlated and if they are we can incorporate the search trend data of neighbouring states into the neural network. Although, we might not be able to perform this task for all the states, we would like to look at some states which initially had been hot spots for the covid-19 spread or were neighbouring such states. We expect there to be some time lag in both the google search trends and covid cases between the hot spots and neighbouring states and we will try to predict this using our model. Lastly, if time permits, we want to explore the effect of media on google search trends and how google search trends of symptoms can be over represented i.e, people experiencing little/no symptoms also search for symptoms because of trends of media platforms.

4 APPROACH

The problem as formulated in the previous section is solved in the same order. The overview of which is shown as follows.

4.1 Study of correlations:

Examining the relationship between Google Trends and the occurrence of COVID-19 is the first step in determining the usability of Google Trends data in the predictability of COVID-19. The Pearson correlation coefficients (r) between the ratio (COVID19 deaths)/(COVID-19 cases) and Google Trends data are calculated since Pearson correlation analysis is the benchmark analysis in this kind of methodology.

$$Pearson's\ correlation\ coefficient(\rho) = covariance(X,Y)/(stdv(X) * stdv(Y))$$
$$cov(X,Y) = (sum(x - mean(X)) * (y - mean(Y))) * 1/(n-1)$$

Next up, the strength of dependency between two variables is then determined using a non-parametric test called the Kendall rank correlation in a secondary correlation study. Te Kendall rank correlation is thought to be reliable for ratio data and is not affected by distribution.

$$tau_b = (P - Q)/sqrt((P + Q + T) * (P + Q + U))$$
$$tau_c = 2(P - Q)/(n * *2 * (m-1)/m)$$

The above two correlation study are performed in order to cross validate the results of one with another. Further, based on the above correlation, we perform the EMD analysis to determine the statistical significance of the input data. The benefit of using the EMD method is that it can eliminate noise and non-stationary oscillations that are unrelated to features of the trend (for instance, informative searches and influential search) that do not correspond to the actual disease (Fever, Asthma etc.) or to the related disease (Covid 19). After removal of the lower order IMFs we again perform the correlation study and validate our results from EMD analysis. The best results after this correlation study are used in further prediction studies.

4.2 Prediction based on data from correlations:

In order to predict the number of cases as described in the problem formulation section in the we will be trying a couple of deep learning models which are recommended for time-series based predictions. These may be from RNNs, LSTMs, GRU etc. These models include a variety of appealing characteristics, including the ability to handle temporal relationships in time-series data, distribution-free learning models, and the adaptability of nonlinear feature modeling.

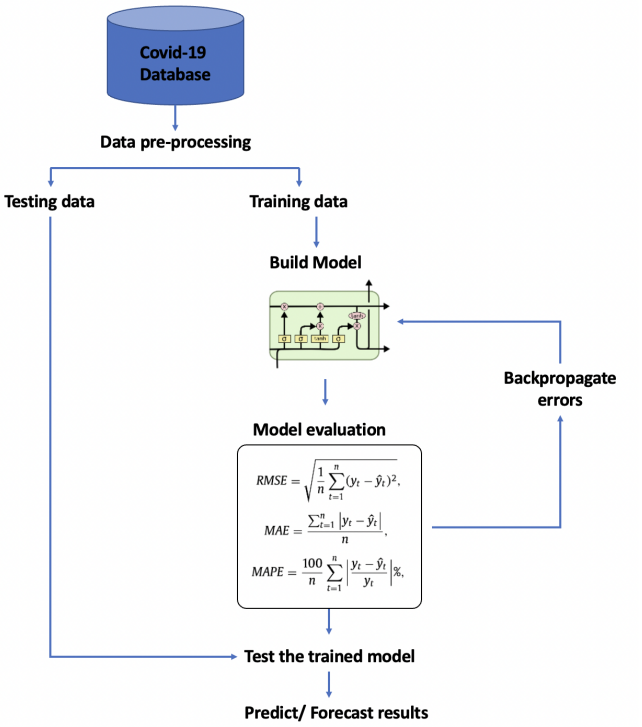

Figure 1: Model pipeline for predicting covid cases based on input data

4.3 Inference:

Inference based on the correlation study from (a) and basic forecasting models from (b) can be made for predicting Covid 19 cases based on the data of symptoms from the neighbouring states of the current state. We first find the correlation between the symptoms of various positively correlated symptoms in their own state with the Covid 19 cases of the current state!

Taking the most correlated symptoms, we will try the deep learning model which we developed in the previous sub-section. If needed we will modify the architecture of our previous Deep Learning model in order to better fit the plots.

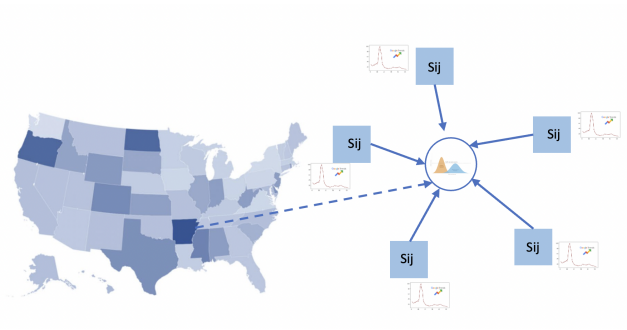

Figure 2: Prediction of Covid 19 cases from trends of neighbouring states

4.4 Model improvement

Finally, we expect to see some errors in the predictions which may corresponds to the fraction of trends that are not due to symptomatic searches. In order to account that, we firstly take one such case (for e.g media coverage) and correlate the data based on media coverage with trends in symptoms. Subsequently we make adjustments in the trends data and use the same pipeline mentioned above to again predict the number of Covid 19 cases with this updated input trend data.

5 PRELIMINARY ANALYSIS OF THE MODEL ON GOOGLE TRENDS DATA

For the proposal we have tried to simulate a basic LSTM model for predicting covid 19 cases in United States based on Google trends data of keyword "Covid". The model seems to match the normalised plots of Covid 19 daily cases. The prelimenary analysis shows decent results due to the good correlation between GT trends and original Covid 19 cases. Articulating the model, we first split our data into testing and training data with a split of 0.58. Since this is a time-series data, hence the split is uniform and continuous and not random. Moving forward, the following architecture has been used for this study. Our current model is working with 2 features in each input, i.e the LSTM is taking in previous 1 week data along with the current week's data as the input to the model.Hence, LSTM in this model is taking input in the form [26,1,2] i.e [samples, time-steps, features]. (Our initial input was a 52 week time-series record of GT trends for keyword: 'Covid' in US)

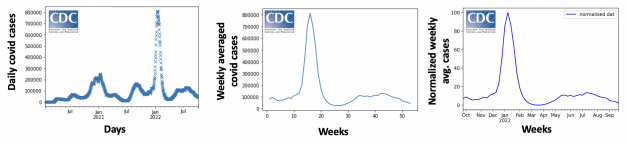

Figure 3: Number of daily Covid case data obtained from CDCs website.

After taking care of the CDC data, now we try to model the system with our LSTM model.

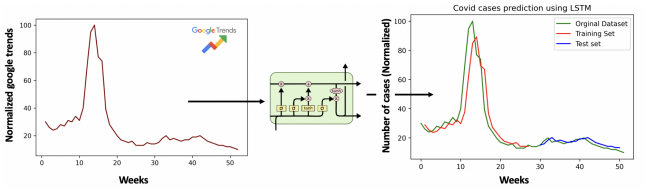

Figure 4: Prediction from a basic LSTM model.

6 CONCLUSION

We have briefly explained how we aim to find correlations between google search trends of various symptoms and covid-19, and how we can further use these correlations to train neural networks and forecast covid-19 cases. Our first step is to verify the correlation between google search trends of symptoms most related with covid-19 and covid-19 cases and then we can explore a different machine learning based approaches for forecasting covid-19 cases.

REFERENCES

[1] Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B Aditya Prakash. 2019. Epideep: Exploiting embeddings for epidemic forecasting. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 577–586.

[2] Kim Asseo, Fabrizio Fierro, Yuli Slavutsky, Johannes Frasnelli, and Masha Y Niv. 2020. Tracking COVID-19 using taste and smell loss Google searches is not a reliable strategy. *Sci. Rep.* 10, 1 (Nov. 2020), 20527.

[3] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences* 454, 1971 (1998), 903–995.

[4] Atina Husnayain, Anis Fuad, and Emily Chia-Yu Su. 2020. Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan. *International Journal of Infectious Diseases* 95 (2020), 221–223.

[5] Tichakunda Mangono, Peter Smittenaar, Yael Caplan, Vincent S Huang, Staci Sutermaster, Hannah Kemp, and Sema K Sgaier. 2021. Information-seeking patterns during the COVID-19 pandemic across the United States: Longitudinal analysis of Google Trends data. *J. Med. Internet Res.* 23, 5 (May 2021), e22933.

[6] Amaryllis Mavragani and Konstantinos Gkillas. 2020. COVID-19 predictability in the United States using Google Trends time series. *Scientific reports* 10, 1 (2020), 1–12.

[7] Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Waqar Aslam, and Gyu Sang Choi. 2020. COVID-19 future forecasting using supervised machine learning models. *IEEE access* 8 (2020), 101489–101499.

[8] Zhaohua Wu, Norden E Huang, Steven R Long, and Chung-Kang Peng. 2007. On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of Sciences* 104, 38 (2007), 14889–14894.

[9] Albert C Yang, Norden E Huang, Chung-Kang Peng, and Shih-Jen Tsai. 2010. Do seasons have an influence on the incidence of depression? The use of an internet search engine query data as a proxy of human affect. *PloS one* 5, 10 (2010), e13728.

[10] Abdelhafid Zeroual, Fouzi Harrou, Abdelkader Dairi, and Ying Sun. 2020. Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos, Solitons & Fractals* 140 (2020), 110121.

Study of inter-regional relations for various disease symptoms generated by Search Trends

Project Milestone Report 1

Anshuman Sinha
Georgia Institute of Technology
Atlanta, US

*Arvind Bangaru (Not contributed in this report)
Georgia Institute of Technology
Atlanta, US

Bhavay Aggarwal
Georgia Institute of Technology
Atlanta, US

ABSTRACT

The use of hospital resources and the development of management plans to best manage infected patients depend on accurate forecasting of COVID-19 cases. However, monitoring sensors is an excellent method to measure the spread of the disease; it is an expensive task and has various privacy and ethical issues. A low-cost alternative to sensors in monitoring the sensors can be leveraging search trends on disease symptoms. It is not far-fetched to think that people with such symptoms would google them to get more information and potential remedies or cures. If this assumption is experimentally validated, it becomes possible to forecast the disease by forecasting symptoms related to Covid-19. In this work, we look at how the disease forecasting problem has been previously approached and how search trends data can be incorporated into it. We also look at predicting pharmacological demands such as vaccines, hospital beds, and medical requirements. Moreover, we also look at the possibilities of including misinformation from media on search trends in order to improve our model.

KEYWORDS

Covid-19, Google search trends, Epidemiology, Deep learning, Time-series, Disease forecasting

ACM Reference Format:
Anshuman Sinha, *Arvind Bangaru (Not contributed in this report), and Bhavay Aggarwal. 2022. Study of inter-regional relations for various disease symptoms generated by Search Trends: Project Milestone Report 1. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 WORK DISTRIBUTION

- Conceptualisation: Equal btw Anshuman S. and Bhavay A.
- Writing: Equal between Anshuman S and Bhavay A.
- Data Collection: Equal between Anshuman S and Bhavay A.
- Work: Initial Findings and Statistical summary, Trends analysis, IMF analysis: Bhavay A.
- Work: Embedding generation and Random projection, Lead time analysis, Spatio-temporal study: Anshuman S.

2 INTRODUCTION

The prediction, modelling and containment of the spread of diseases is one of the most important and challenging problems of the modern world. In our previously submitted project proposal, we formulated our problem of 'Studying the inter-regional relations for various disease symptoms generated by search trends data in the context of Covid-19'. The overall study is divided into the following sub-parts:

- **Studying correlations:** Finding the top-k most correlated symptoms from our trends data. (details of the process given in section 6 and 7) And further expanding the study to inter-regional correlation of symptoms and cases.
- **Inference study:** Based on the study of co-relation we observe the various features of the pandemic such as:
  - i **Phase changes** and its spatial correlation between neighbouring states. We will study and compare the trends in search data and Covid cases; And also study the effect of changing trends on neighbouring states.
  - ii Study of **lead time** and inferring how search data from a better connected state can be used to restrict surge in cases of the neighbouring states.And extending this to find lead time over surge in vaccine demands of the state.
  - iii Correlation between search trends and **targets** like, hospitalisation, severe infections, mortality etc.
- **Predictions:** In this work, we will perform Covid-19 prediction in the following cases.
  - i Prediction of Covid-19 cases with the help of trends of most correlated symptoms of the state and extending it to prediction based on neighbouring states.
  - ii Comparing the results of lead time with our predicted cases and providing a validation to our model.
  - iii Predicting the demand of pharmacological amenities such as vaccines, drugs based on search trends; such that the disease can be better managed.
- **Model improvement:** Study the impact of media influence (such as twitter trends) on search trend's shortcoming of over-predict the related diseases and eventually its effects on estimates of Covid-19 (In sense, studying the impact of miss-information).

We will be working on the above mentioned topics following the same order as listed above. As of now, we have started working on Inference study (finished most of the parts, results shown later in the draft).

3 RESPONSE TO INITIAL COMMENTS

- ***'Try to get papers from well known journals and improve the quality of work'***: We have read and implemented

Conference'17, July 2017, Washington, DC, USA
2022. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

research papers from better journals and conferences in various sections of our work. Such as exploring the possibilities of using other data-sets[13] [6] [10] , Implementing algorithms to better our current model [12] [11] [3], Improving performance of our current model by implementing better feature engineering.[2] [8]

- ***'Make sure you compute correlations and predictions in a real-time..'***: This is what we had initially planned to do but did not mention it explicitly in our proposal. Our LSTM model defines a time-steps that the model we look in the past while predicting any future value. Somewhat like a sliding window over the data-set.
- ***'Project needs to be expanded a bit..'***: We had initially planned on predicting Covid cases of a selected state from the Google trends data of its neighboring states and also to study the effect of miss information like media trends affect the trends in symptoms search data. Now, we will be extending our work to
  i Trends study: Phase changes and its effects.
  ii Lead time analysis: Predicting lead time through search trends.
  iii Targets study: Hospitalisation, Vaccine requirements etc.
  iv Anomaly detection: Anomalous data based on trends of neighbouring states.
  v Spatio-temporal study: Studying the spatial effect of all the above mentioned points.
- ***'There might be a CS problem here on finding top-k correlated time-series efficiently'***: Although there can be many approach towards solving this problem, we have decided not use any predefined library but to write our own method. In this work are using Random projection method as mentioned in the Query stream search paper by Chien et al. [2] (Which was mentioned in our comments)

4 REVIEW OF OUR PREVIOUS WORK

In the proposal submission, we had formulated the problem and defined our approach towards the solution of our problem (Refer Problem formulation and Approach section of Proposal).

We had also simulated a basic LSTM model for predicting Covid 19 cases in United States based on Google trends data of keyword "Covid" from the start of year 2022 to 50 weeks into the year. The model seems to match the trends plots of Covid 19 daily cases. The preliminary analysis shows decent results due to the good correlation between GT trends and original Covid 19 cases. Figure 1 and 2 shows our work from the initial model.

5 DATA COLLECTION APPROACH

In the current work will be implementing data from various search sources mentioned in table 1. These datasets are implemented according to the model, such as trends data from Google and Facebook are used as symptomatic predictions, Hospital capacity data for studying the measure of targets , The media coverage data in improving the model etc. The weekly covid cases data was downloaded from CDC. Daily google trends data was downloaded from 2020 to 2022 and resampled into a weekly format. As detailed by the

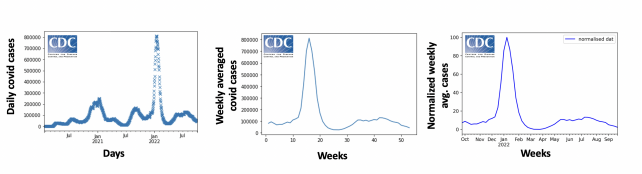

Figure 1: Number of daily Covid case data obtained from CDCs website.

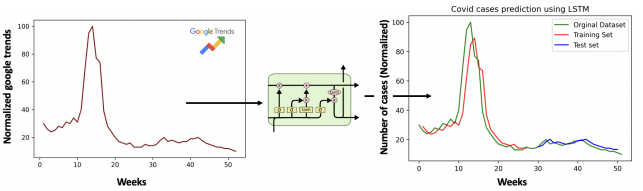

Figure 2: Prediction from a basic LSTM model.

authors in [1], the search trend is defined as the ratio between the normalized search volume for a given symptom in a given region during a given week and the median search volume for that region-symptom pair. This is done to ensure that the google search are comparable between regions.The timeframe of the data was restricted to start from $2020-01-26$ and end at $2022-10-13$. These dates coincide with the first week of covid cases and the last week of reported covid cases available on CDC and consists of a total of 141 weeks. We also used the daily cases in prediction of lead time, by converting the data to lower a dimension first and then applying the later algorithms. We plan to begin incorporating the other mentioned datasets in the future.

6 DESCRIPTION OF INITIAL FINDING AND STATISTICAL SUMMARY OF DATASET:

Examining the relationship between Google Trends and the occurrence of COVID-19 is the first step in determining the usability of Google Trends data in the predictability of COVID-19. The Pearson correlation coefficients (r) between the ratio (COVID19 deaths)/(COVID-19 cases) and Google Trends data are calculated since Pearson correlation analysis is the benchmark analysis in this kind of methodology. We perform our experiments on the national level and also for the states - California, New York, Texas, Alaska, Mississippi and Georgia.

We first checked the PCC (similar to the work of [9] and [5]) of the weekly covid cases with the search trends for all the symptoms. The top-10 symtoms and the corresponding PCC values are listed in table 2 This approach however has a major drawback, it assumes that the variance in the data is homogeneous across the data range.

We observed that many unrelated symptoms ended up having a high correlation value. Although, these symptoms can be manually filtered, we wanted to look at how correlated the symptoms in smaller windows over the entire time frame. We created rolling windows of 4 weeks for both the datasets and performed Pearson's correlation on the symptom window with the corresponding cases

Application	Data type	Publisher	Link
Covid 19 cases data	COVID-19 statistics	Johns Hopkins University	github/CSSEGISandData/COVID-19
Covid 19 cases data	COVID-19 statistics	CDC	cdc/2019-ncov
Measuring emotions	Textual/Embeded data	COVID-19 Real World Worry	github/ben-aaron188/covid19worry
Media Coverage	National and International news article and interpretation	Humanities and social sciences communication [nature]	nature/News-media-coverage-of-COVID-19
Search trends data	Normalised keyword search data	Google	google/search-trends
COVID-19 Trends and Impact Survey	Symptoms and behaviour data	Facebook and Delphi CMU	cmu-delphi/symptom-survey
Hospital Capacity Data	Number of beds occupied and available state wise data	U.S. Department of Health & Human Services	healthdata/capacity-bed-data

Table 1: Data Sources

Symptom	PCC
Hypoxemia	0.643161
Eye pain	0.635909
Night sweats	0.624184
Nasal congestion	0.582434
Sinusitis	0.545885
Bradycardia	0.525026
Throat irritation	0.515732
Ageusia	0.500023
Post-nasal drip	0.491759
Anosmia	0.483690

Table 2: Global Person's Correlation between symptoms and covid cases

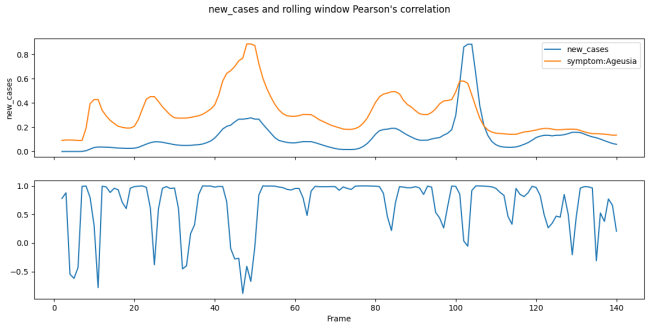

Figure 3: Plot of local window search trends of Ageusia along with the weekly covid cases for the entire nation. The bottom plot shows the PCC for the rolling window.

window (table 3). The table shows the symptoms which appeared the most in the top-15 most correlated symptoms. 50% indicates the median PCC value across all windows and is the metric we focus on(symptom with maximum value highlighted for each region). For the selected regions, Ageusia has the highest median PCC value majority of the times, so it is a symptom we can explore a bit more. The disease "ageusia" refers to the loss of sense of taste and is prominent symptom among people infected with covid. Moreover, the other top symptoms are also prominent symptoms of covid. This result shows the rolling window/local PCC is an effective method to extract the most correlated symptoms.

Next up, we perform the EMD analysis as decribed by [14], [4] and [7] to determine the statistical significance of the input data. The benefit of using the EMD method is that it can eliminate noise and non-stationary oscillations that are unrelated to features of the trend (for instance, informative searches and influential search) that do not correspond to the actual disease (Fever, Asthma etc.) or to the related disease (Covid 19). We removed the lower order IMF's and combined the first and second order IMF to form the new curve for each symptom's google trends. We again performed local PCC on the new symptom trend curves with the covid cases. We compared the correlation obtained this way with the original local PCC and observed that the correlation was comparatively lower. This suggests that the assumption we made that the oscillations were noise in the trends data was incorrect and these oscillations are important trends. Our next step to limit the overrepresentation of symptom search trends induced by media/social media would be to incorporate the media coverage data listed in table 1.

7 DESCRIPTION OF MATHEMATICAL BACKGROUND NECESSARY FOR THE PROBLEM:

A background knowledge of Statistics, Probability and Linear algebra is required to understand the working of our models. We briefly point the highlights of the fundamentals which we have used in our current work:

7.1 Random projections

Since, our time-series data was of very high dimension (e.g. Dim 52 for Weekly data and Dim 365 for Daily data) we are required to lower the dimension of the data as well as transform it such that it takes less space.

		Low-grade fever	Fever	Pneumonia	Ageusia	Hypoxemia	Dysgeusia	Anosmia	Shortness of breath
National	mean	0.55	0.47	0.41	0.65	0.56	0.49	0.46	0.36
	25%	0.38	0.02	-0.01	0.46	0.22	0.19	0.00	-0.02
	50%	0.86	0.82	0.76	**0.92**	0.85	0.73	0.72	0.51
	75%	0.96	0.96	0.95	0.98	0.97	0.93	0.97	0.93
California	mean	0.42	0.39	0.28	0.57	0.52	0.43	0.40	0.30
	25%	0.01	-0.02	-0.32	0.37	0.28	0.02	-0.02	-0.11
	50%	0.62	0.61	0.52	**0.81**	**0.81**	0.62	0.61	0.42
	75%	0.95	0.92	0.91	0.96	0.96	0.93	0.92	0.89
Texas	mean	0.35	0.30	0.45	0.45	0.49	0.36	0.28	0.30
	25%	-0.07	-0.17	0.12	0.12	0.00	0.00	-0.23	-0.12
	50%	0.57	0.57	0.68	0.68	**0.76**	0.54	0.48	0.49
	75%	0.87	0.89	0.94	0.94	0.94	0.87	0.89	0.86
New York	mean	0.62	0.49	0.35	0.62	0.62	0.42	0.50	0.43
	25%	0.38	0.17	-0.03	0.40	0.38	0.03	0.23	0.05
	50%	0.85	0.76	0.65	**0.86**	0.85	0.66	0.73	0.65
	75%	0.95	0.93	0.93	0.97	0.95	0.88	0.94	0.91
Alaska	mean	0.13	0.31	0.26	N/A	N/A	N/A	N/A	0.14
	25%	-0.58	-0.06	-0.18	N/A	N/A	N/A	N/A	-0.31
	50%	0.32	**0.57**	0.35	N/A	N/A	N/A	N/A	0.24
	75%	0.80	0.84	0.84	N/A	N/A	N/A	N/A	0.60
Mississippi	mean	0.33	0.36	0.21	0.24	0.30	0.16	0.25	N/A
	25%	0.00	-0.01	-0.28	-0.25	-0.12	-0.30	-0.24	N/A
	50%	0.48	**0.62**	0.26	0.43	0.51	0.21	0.40	N/A
	75%	0.87	0.91	0.80	0.86	0.87	0.67	0.86	N/A
Georgia	mean	0.45	0.40	0.33	0.53	0.50	0.33	0.46	N/A
	25%	0.01	-0.04	-0.08	0.18	0.02	-0.03	0.02	N/A
	50%	0.76	0.70	0.62	0.71	**0.81**	0.43	0.70	N/A
	75%	0.94	0.94	0.92	0.97	0.97	0.85	0.92	N/A

Table 3: Local Person's Correlation between symptoms and covid cases

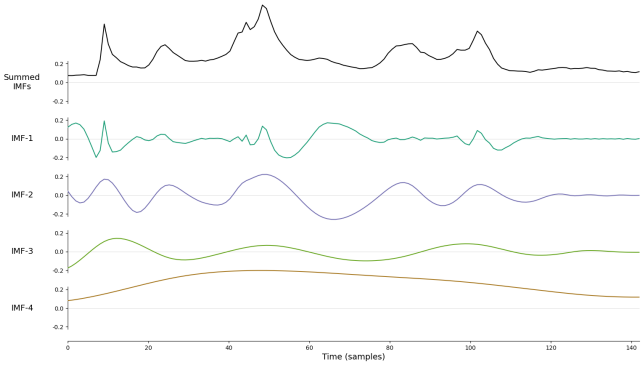

Figure 4: Google search trends for "Ageusia" divided into its IMF components

Random projection is used to project the time-series vector on a random hyperplanes, such that we can evaluate which side of the hyperplane these vectors are lying. Further, we randomly pick a sufficient number of hyperplanes from a Gaussian distribution to remove the stochasticness involved in the process. Figure 4, shows the schematic of this process.

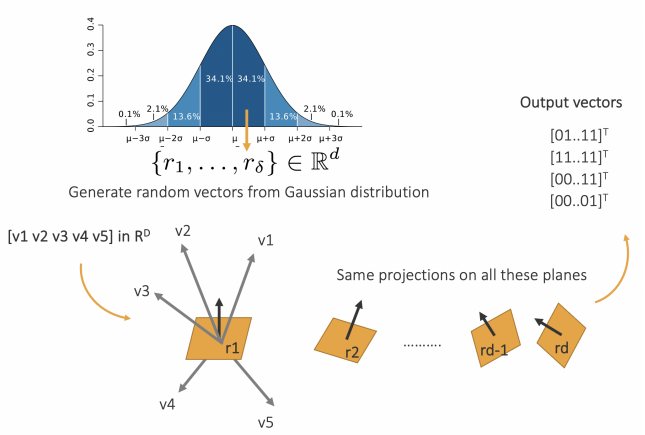

Figure 5: A pictorial representation of the main question that the Johnson-Lindenstrauss Lemma resolves.

7.2 Johnson-Lindenstrauss Lemma

In order to understand that vector representation on the low dimensional subspace (or on the hyperplane) is similar to the original distance between these two vectors we need the Johnson-Lindenstrauss Lemma.

Informally, the lemma states that we can map N points into a much smaller Euclidean space, specifically one of logarithmic dimension, while maintaining the pairwise Euclidean distances of the given points up to a multiplicative factor of 1 +. N points are defined as being in the N-dimensional Euclidean space.

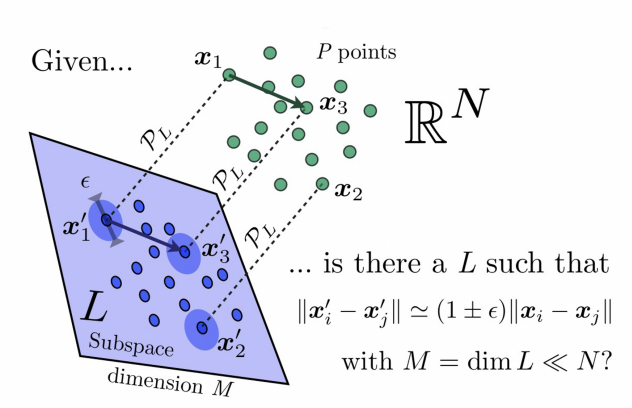

Figure 6: A pictorial representation of the main question that the Johnson-Lindenstrauss Lemma resolves.

7.3 Statistical correlation

In order to find how similar the google search trends is similar to the weekly covid cases, we are using Pearson's correlation. Pearson's correlation measures the statistical relationship between two continuous variables and outputs a number between −1 and 1 that measures the strength and direction of the relationship between two variables.

$$Pearson's\ correlation\ coefficient(\rho) = covariance(X,Y)/(stdv(X) * stdv(Y))$$

$$covariance(X,Y) = (sum(x - mean(X)) * (y - mean(Y))) * 1/(n-1)$$

8 FORMAL DESCRIPTION OF IMPORTANT ALGORITHMS USED

8.1 Binary Embedding using LSH

- Make k random vectors of length d each, where d is the feature vector's dimension and k is the size of the bitwise hash value.
- Calculate the dot product of the random vector and the observation for each random vector. If the dot product result is positive, set the bit value to 1; otherwise, set it to 0.
- Concatenate all of the bit values that were calculated for the k-dot products.

8.2 Spatial effects on lead time

- Find the top-k most correlated symptoms of a state.
- Find the top-k most correlated symptoms of the neighbouring states.
- Compute the lead time of each neighbouring state from their symptoms.
- Re-compute the lead time of each neighboring state from the symptom of the original state.
- Evaluate the gain in lead time for each neighbouring state.

National data	Top k symptoms	Match	Lead time (t*)
1	Anosmia	18	271
2	Acute bronchitis	17	267
3	Ageusia	17	271
4	Bronchitis	17	267
5	Chills	17	271
6	Common cold	17	267
7	Cough	17	267
8	Dysgeusia	17	270
9	Fever	17	14
10	Hemoptysis	17	270
11	Hypoxemia	17	271
12	Low-grade fever	17	7
13	Nasal congestion	17	271
14	Pneumonia	17	6
15	Post-nasal drip	17	15

Table 4: Lead Time of symptom search trends on national weekly covid cases.

9 SOME INITIAL RESULTS

9.1 Lead time analysis

The lead time analysis was performed on the National covid (daily cases) data with respect to the symptoms data (for year 2022, Jan to Oct; 274 dimesnion vector). We first found the most correlated symptom with the help of top-15 symptom search (using dimension reduction) and then analyse the lead time on the original data wrt. to these most correlated symptoms. Table 4 shows the result which shows logical lead time wrt. some important symptoms like 'Fever', 'Low-grade Fever' etc and many of these search trends tend not to align well with the original data (hence, we get very high lead times)

9.2 Spatio-temporal study

We also studied the most correlated symptom of neighbouring states and whether they give us better lead time than the original state or not! For this study we took neighbours of Georgia (Alabama and South Carolina) and studied the lead time of these neighbouring states wrt. the symptoms searches in Georgia. The results are detailed in tables 5 and 6 for Alabama and South Carolina respectively.

Here, we observe that we get a better lead time for the neighbors with the most correlated symptoms of Georgia than from their own symptoms search trend. These results show the possibility of spread of covid from Georgia in these states, and also tells that we can get a better insight (lead time) with a neighbouring state (which maybe better connected to other states of the country) and make use of

Top k symptoms	Match	t^*	t^* wrt Georgia	Gain in t^*
Ageusia	27	10	11	1
Hypoxemia	27	8	8	0
Shortness of breath	27	49	47	-2
Anosmia	26	4	10	6
Bradycardia	26	49	49	0
Erectile dysfunction	26	49	49	0
Tachycardia	26	49	49	0
Chest pain	25	47	49	2
Halitosis	25	44	44	0
Headache	25	13	15	2
Middle back pain	25	49	49	0
Pruritus ani	25	47	49	2
Dysgeusia	24	10	11	1
Hair loss	24	48	49	1
Hypoxia	24	49	49	0

Table 5: Lead Time of Georgia's symptom search trends on weekly covid cases in Alabama.

Top k symptoms	Match	t^*	t^* wrt Georgia	Gain in t^*
Chest pain	25	49	49	0
Shortness of breath	25	12	47	35
Chills	24	0	0	0
Dysgeusia	24	9	13	4
Hypoxemia	24	7	9	2
Ageusia	23	10	15	5
Bradycardia	23	47	49	2
Erectile dysfunction	23	49	49	0
Eye pain	23	48	11	-37
Low-grade fever	23	0	12	12
Night sweats	23	47	47	0
Tachycardia	23	49	49	0
Anosmia	22	5	10	5
Fever	22	5	13	8
Hemoptysis	22	48	1	-47

Table 6: Lead Time of Georgia's symptom search trends on weekly covid cases in South Carolina.

this gain in lead time to address the forthcoming situation in the state.

10 GENERAL DIFFICULTIES IN OUR WORK

- Top k searches : We were getting very high dimensional data (especially when considering daily cases).
- Data collection : The data sources suitable for our work was initially difficult to find (Some had permission issues, e.g. AWS data)
- Anomaly prediction not easily quantifiable(as time-series is limited) : In our earlier approach, we found that the dataset for Covid is still not having sufficient amount of seasons to detect anomoly. But now we are using anomaly prediction in a more general sense wrt. to neighbouring states symptoms.

11 FUTURE WORK

- We have explored whether google search trends are correlated with covid cases on different spatial levels, and we now aim at including the other trends data sets in the upcoming weeks.
- We have found several correlated symptoms which also match with the covid-19 symptoms listed by the CDC. We have utilized these symptoms to study lead time at a national level and between neighboring states.
- Moving forward, we want to consolidate our top-k correlated symptoms and perform our subsequent experiments using them. We will further expand our spatio-temporal study of the impact of neighboring states google trends to study lead time.
- We have to also begin to work on forecasting models.
- Finally we also have to address model improvement (as described in the section 'Introduction')

More detail can be found on the Introduction section.

REFERENCES

[1] Shailesh Bavadekar, Andrew Dai, and Davis et al. 2020. Google COVID-19 Search Trends Symptoms Dataset: Anonymization Process Description (version 1.0). https://doi.org/10.48550/ARXIV.2009.01265

[2] Steve Chien and Nicole Immorlica. 2005. Semantic similarity between search engine queries using temporal correlation. In *Proceedings of the 14th international conference on World Wide Web*. 2–11.

[3] Vinay Kumar Reddy Chimmula and Lei Zhang. 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals* 135 (2020), 109864.

[4] Ramon Gomes Da Silva, Matheus Henrique Dal Molin Ribeiro, Viviana Cocco Mariani, and Leandro dos Santos Coelho. 2020. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. *Chaos, Solitons & Fractals* 139 (2020), 110027.

[5] Jacques Demongeot, Yannis Flet-Berliac, and Hervé Seligmann. 2020. Temperature decreases spread parameters of the new Covid-19 case dynamics. *Biology* 9, 5 (2020), 94.

[6] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.

[7] Najmul Hasan. 2020. A methodological approach for predicting COVID-19 epidemic using EEMD-ANN hybrid model. *Internet of Things* 11 (2020), 100228.

[8] Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. 604–613.

[9] Amaryllis Mavragani et al. 2020. Tracking COVID-19 in Europe: infodemiology approach. *JMIR public health and surveillance* 6, 2 (2020), e18941.

[10] William JM Probert, Chris P Jewell, Marleen Werkman, Christopher J Fonnesbeck, Yoshitaka Goto, Michael C Runge, Satoshi Sekiguchi, Katriona Shea, Matt J Keeling, Matthew J Ferrari, et al. 2018. Real-time decision-making during emergency disease outbreaks. *PLoS computational biology* 14, 7 (2018), e1006202.

[11] Siva R Venna, Amirhossein Tavanaei, Raju N Gottumukkala, Vijay V Raghavan, Anthony S Maida, and Stephen Nichols. 2018. A novel data-driven model for real-time influenza forecasting. *Ieee Access* 7 (2018), 7691–7701.

[12] Svitlana Volkova, Ellyn Ayton, Katherine Porterfield, and Courtney D Corley. 2017. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS one* 12, 12 (2017), e0188941.

[13] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1

[14] Albert C Yang, Jong-Ling Fuh, Norden E Huang, Ben-Chang Shia, Chung-Kang Peng, and Shuu-Jiun Wang. 2011. Temporal associations between weather and headache: analysis by empirical mode decomposition. *PloS one* 6, 1 (2011), e14612.

Study of inter-regional relations for various disease symptoms generated by Search Trends

Project Final Report

Anshuman Sinha
Georgia Institute of Technology
Atlanta, US

Arvind Bangaru
Georgia Institute of Technology
Atlanta, US

Bhavay Aggarwal
Georgia Institute of Technology
Atlanta, US

ABSTRACT

The use of hospital resources and the development of management plans to best manage infected patients depend on accurate forecasting of COVID-19 cases. However, monitoring sensors is an excellent method to measure the spread of the disease; it is an expensive task and has various privacy and ethical issues. A low-cost alternative to sensors in monitoring the sensors can be leveraging search trends on disease symptoms. It is not far-fetched to think that people with such symptoms would google them to get more information and potential remedies or cures. If this assumption is experimentally validated, it becomes possible to forecast the disease by forecasting symptoms related to Covid-19. In this work, we look at how the disease forecasting problem has been previously approached and how search trends data can be incorporated into it. We have explored the possibility of using search trends as a proxy to actual disease cases with the help of SIR-Network model. We extend the prediction of cases to inter-regional space as well, with spatio-temporal predictions with the help of search trends. We also looked at predicting pharmacological demands such as vaccines, and miss-information of disease through search trend results..

KEYWORDS

Covid-19, Google search trends, Epidemiology, Deep learning, Time-series, Disease forecasting, Pytorch-forceast, SIR-Network

ACM Reference Format:
Anshuman Sinha, Arvind Bangaru, and Bhavay Aggarwal. 2022. Study of inter-regional relations for various disease symptoms generated by Search Trends: Project Final Report. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 WORK DISTRIBUTION

- Work: Correlation study, Spatio-temporal study, IMF component study, Deep learning based predicitons : Bhavay A.
- Work: Population netowork creation : Arvind B.
- Work: Lead time study, Spatio-temporal study, Targets study, SIR-Network analysis : Anshuman S.

2 INTRODUCTION

Covid-19 is a viral contagious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus. It has spread globally and was declared a global pandemic in march 2020. Covid-19 is a flu-like disease and has similar methods of transmission. Symptoms of covid-19 Include fever or chills, cough, shortness of breath or difficulty breathing, fatigue, muscle or body aches, headache, new loss of taste or smell, sore throat, congestion, nausea, and diarrhoea among others. Managing the pandemic involves collecting data in forms of surveys, Covid-19 test results and electronic health records. There are challenges to collecting and using these data like large-scale surveying being expensive, time-consuming, and having limitations in terms of capturing time-sensitive data and accuracy. Electronic health data requires complex anonymization, merging and analysis before they can be used. Due to these limitations, alternative approaches like information-seeking behaviours of the population captured in data like Google search trends are an exciting avenue for forecasting and managing pandemics. Covid-19 has had far-reaching economic, cultural, and social impacts and has led to the COVID-19 recession. Predicting and planning for local covid-19 outbreaks is very important from a public health perspective and this study helps in this prediction from publicly available data. The prediction, modelling and containment of the spread of diseases is one of the most important and challenging problems of the modern world!

Conference'17, July 2017, Washington, DC, USA
2022. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 PROBLEM DEFINITION

The prediction, modelling and containment of the spread of diseases is one of the most important and challenging problems of the modern world. The overall study is divided into the following sub-parts, we have tried to study each of these problems in a chronological fashion to achieve end-to-end learning:

- **Studying correlations:** Finding the top-k most correlated symptoms search trends to the covid cases. We further study to inter-regional correlation of symptoms and cases.
- **Inference study:** Based on the study of correlation, we observe the various features of the pandemic such as:
  - i **Phase changes** and its spatial correlation between neighbouring states. We will study and compare the trends in search data and covid cases; And also study the effect of changing trends on neighbouring states.
  - ii Study of **lead time** and inferring how search data from a better-connected state can be used to restrict surge in cases of the neighbouring states. And extending this to find lead time over the surge in vaccine demands.
  - iii Correlation between search trends and **targets** like, vaccination, miss-information .
- **Forecasting:** In this section, we performed covid-19 forecasting using deep learning models in the following cases.

i Prediction of covid-19 cases with the help of trends of most correlated symptoms of the state.
ii Using the most correlated symptoms of a given state to predict the covid-19 cases of its neighbouring states.
iii Comparing the results of lead time with our predicted cases and providing validation to our model.

- **Search trends based SIR-Network model:** Predicting Covid-19 cases based on search trends with mechanistic models.
  i Using symptom search trends data as proxy to covid 19 cases; in order to estimate the time-varying parameters of mechanistic models.
  ii Build contact networks from synthetic contacts datasets at county level and simulate SIR mode of spread on the network with the evaluated model parameters.
  iii Compare the trends in cases obtained by SIR-network based model with actual cases to check the efficiency.

4 RELATED WORK AND SURVEY

Several studies have been conducted using correlation and regressing models [2, 7, 8, 10, 12] to model and forecast Covid-19 cases using Google Trends. They could further improve the models using strategies such as using quantile regression model for forecasting [12], correcting for extraneous influences like media coverage of those specific symptoms[2, 8] and lead time analysis [8]. Other extensions to forecasting include using Google Trends as a proxy for what people are thinking, needing, and planning in real-time to understand population behaviour, public health decision making [10]. For seasonal disorders like depression, The empirical mode decomposition (EMD) as sued to break down complex time series into a set of intrinsic oscillations, which were used for correlation and forecasting [7]. These studies show that Google Trends and COVID-19 data display statistically significant relationships and the models show good COVID-19 predictability after they have been corrected for some extraneous influences.
Mechanistic models try to mathematically model the spread of infection in communities. Agent-based network models are a kind of mechanistic models which try to represent each person in the community as a node in a contact network. Studies have used poissonian Small-World Network with transmission rate (beta) estimated from COVID-19 individual-level secondary attack rate to expain COVID-19 infection curves [16]. They also showed that such approach of parameter estimation worked even when used on large scale networks and used this to explain real covid-19 infection curves. Recent works such as GRADABM [3] have used stochastic disease transmission model to model tensor-based agents to produce gradient estimates for disease spread and speed up network-based models. Constructing networks for Agent-based mmodels is a challenge. One way to overcome this is to use synthetic contacts networks such as the one by biocomplexity institute at the University of Virginia [13] which contains mixing matrices for the current synthetic U.S. population and contact networks constructed from it.

Forecasting COVID-19 cases using machine learning The use of hospital resources and the development of management plans to best manage infected patients are dependent on accurate forecasting of the number of COVID-19 cases. The past couple of years has seen a great rise in the number of studies exploring machine learning and deep learning for this use. Comprehensive comparisons of Machine learning models for prediction of COVID-19 cases [14] and deep learning models for forecasting the time-series data of COVID-19 [18] reveal that Variational AutoEncoder (VAE) performed the best and were closely followed by Long Short-Term Memory (LSTM) which address the vanishing gradient issues that restrict the effectiveness of a straightforward Recurrent Neural Networks (RNN). Other studies for infuenza forecasting have used Improved Deep Embedded clustering (IDEC) to learn from historical influenza data and then used LSTM based encoder-decoder module to make valuable predictions such as onset week, seasonal peak week, seasonal peak intensity and future incident for influenza after correcting using an attention mechanism [1]. The enhanced capabilities of deep learning models to capture process non-linearity and their adaptability in modeling time-dependent data are major driving forces behind these applications.

5 PROPOSED METHOD

Intuition: Google search trends serve as indicators of the population's overall interest. Studying trends can give us an idea of what events are important and what interests the population. Extending this to study disease spread, the trending symptoms can give us a potential indication of the symptoms of the spreading disease. Using this for the COVID-19 pandemic, we can track the spread of the disease as the rise in search trends in theory correlated to the rise in cases we can expect people to search more about the symptoms they or the people near the are facing. By then utilizing the most correlated symptoms, we can try to forecast the spread of the disease, and in this way, we do not have to rely on covid data, which might be unreliable and delayed.

- Our positive preliminary results with basic LSTM, which was able to predict the cases at the national level with the help of search trends only. (Ref: Initial submission)
- Secondly, we are able to get a good correlation between neighbouring states' covid cases and the present state's most correlated symptoms. Which pointed towards further predicting these cases with advanced models (Ref: Prediction).
- According to our review, we did not find any model which was learning the transient parameters of the epidemic through search trends independently.
- Note: Although, due to the efforts by public health agencies and the government; the data related to covid is extensively available. But that's not true for many other diseases.

Description: We have looked at the various aspects of utilising google search trends in predicting and containing the spread of disease and providing use-full information about the trends of the targets (like awareness among public (-ve RT-PCR tests) , shortage of vaccines (vaccination data). In this section we will describe our approach, algorithm and model implementation for our current work.

1. Approach:

- **Defining the outcomes:** As discussed in the problem formulation section, we first defined our end goals.(Ref:Sec 3)

- **Data collection**: The first step to our project was collecting the necessary data for our model. The details of our data sources and their application is shown in Table 5
- **Data pre-processing**: The google search trends have been collected from 2017, but for our current approach, we do not have any way to incorporate this data. Thus, we only include the data from when covid began, which starts from the date $2020-01-26$ and we use the data reported to date $2022-10-13$. In the end, we have weekly search trends and daily covid case data for 42 weeks for each state in the United States. The state data was aggregated to create national data.
- **Modelling decision:** Several decisions had to be taken while considering which models to pick for the analysis of these datasets. These decisions include, Choice of correlation metric, choice of deep learning models, mechanistic models etc.
- **Necessary changes:** With time we had to make a lot of changes in both our end goals and our approach (in terms of models) on the basis of our intermin results. E.g., We decided to encorporate top-search method to improve efficiency of our work, We decide to switch to Pytorch's forecasting models after LSTMs were not giving good results. We moved to Network-SIR model from SI and SIS models.

2. Algorithm:

- **Correlation study**:
  - We first checked the PCC (similar to the work of [11] and [5]) of the weekly covid cases with the search trends for all the symptoms.
  - We also created rolling windows of 4 weeks for both the datasets and performed Pearson's correlation on the symptom window with the corresponding cases window.
- **Top-k correlated symptoms**:
  - With the help of random projections method compute the top most correlated symptoms.
  - The top 15 most occurring symptoms were then selected for further experiments.
- **Lead time study:**
  - With he help of top-k most correlated symptoms we evaluate the lead time of cases with respect to trends.
  - Shift the data of search trend such that maximum correlation is found between search trend and cases.
  - The greatest shift with max correlation is the lead time.
- **Target study:**
  - With the top k-most correlated symptoms, we perform the correlation study on targets. In our present study we have looked at two such targets.
  - Firstly study of vaccine requirements, by studying the correlation between symptoms search trend and vaccination.
- **Network formation**:
  - For each county, we first created nodes equal to the numbers from the synthetic contacts dataset and labeled each node with their age group. The 5 CDC-defined age groups.
  - Then, we independently randomly sampled nodes from two age groups and create a directional edge in between.
  - This is repeated till the total number of edges matched the total contacts between the two age groups in the synthetic contacts dataset.
- **SIR model parameter estimation:**
  - Performing time-varying parameter estimation of SIR model, with the help of google search trends as proxy infection.
  - The model works on a 7-day moving window and minimises the errors produced by SIR's ode solutions and google search trend data.
  - The next 7-day parameter prediction is followed by the previous estimated values as initial conditions to the model.
- **SIR-Network model:**
  - Start the model with help of time-varying SIR parameters and Population network from meta-data.
  - The spread/dynamics of infection changes according to β and γ parameters on the network.
  - We apply this model on Hot Springs County 56017 of Wyoming (with the least population) in the view of computation limits and time.
  - Evaluate the model with Wyoming's CDC data of cases.

3. Model:

- **Time-series model**: We performed forecasting of covid-19 cases using the google search trends after establishing the correlation between the symptom trends and covid cases. For forecasting, we started of with a vanilla LSTM and then we moved on to more advanced models. Due to time and computational constraints, we did not create any advanced model from scratch and instead used the implementations provided by Pytorch Forecasting. We used two such models DeepAR[15] and Temporal Fusion Transformer[9]. DeepAR is a autoregressive recurrent network that can be trained on multiple time series and utilizes quantile loss to make prediction intervals based on its confidence. DeepAR utilizes LSTM's to parameterize the Gaussian Likelihood Function instead of directly using them to make predictions. Temporal Fusion Transformers build upon LSTM's and integrates the attention mechanism to them. The attention mechanism allows the model to learn time-varying relationships and the model also includes gating components which enable skipping of unnecessary parts of the network and pick relevant input features at each time step. Moreover, it also makes the model's predictions more interpretable.
- **Mechanistic model: SIR ODE model**: Mechanistic model The COVID-19-related deaths as well as recovered persons make up the deleted group in our application. The following presumptions will be used to build this model:
  - After healing, people won't be susceptible anymore.
  - The infection will be the only factor in deaths.
  - No new infection will be added to the population.
  - The likelihood of susceptibility, infection, and removal is the same for every person and governed by model parameters.

(1) $S'(t) = \frac{d[S(t)]}{dt} = -(\frac{\beta}{N})(S(t)I(t))$

(2) $I'(t) = \frac{d[I(t)]}{dt} = (\frac{\beta}{N})(S(t)I(t)) - (\gamma)[I(t)]$

(3) $R'(t) = \frac{d[R(t)]}{dt} = - (\gamma)[I(t)]$

Further optimisation of β and γ are computed with the help of training data (Search trend's data). We used search trends as ground truth $I_{search}(0) \ldots, I_{search}(T)$. The objective function to minimize is

$$\mathcal{L} = \sum_{t=1}^{T} (I(t) - I_{search}(t))^2$$

The corresponding parameters are computed with a moving window, such that the time-varying trends in parameters is captured.

- **Agent-based network models: SIR-Network model**
  A SIR model can alternatively be constructed as a network as an alternative to employing differential equations, as in the prior model. Each node in the network here corresponds to a person. A disease can spread through social ties, which are represented by the edges connecting nodes. Neighbors of an infected node are infected with a specific likelihood (Chance infection) in each iteration step, while those who have already contracted the virus recover with a similar probability (Probability recovery). In contrast to the equation-based version, the network's topology has an impact on how the disease spreads. The population map on the right of the network shows this.

6 RESULTS

Questions answered: In this section we are highlighting the most important questions which our current study tries to answer.

(1) Scope of using search trends as in predicting covid-19 cases?
(2) Do we have any correlation between search trends and covid cases of the neighbouring states?
(3) How much lead time of changing epidemic trends can we get using only search trends?
(4) What are the rumours related to covid symptoms?
(5) What are the scope of using mechanistic model based on search trends.

Experiments: In this section, we have listed all the experiments which we have performed in order to answer the above question. Our observations and Findings are sub-listed with each of the experiment.

1. **Correlation between search trend and covid cases:** (Case - US-National, Georgia, California, Texas, New York, Alaska, Mississippi):
   - Observation: Table 1, Table 2, Figure 2
   - Findings:
     The top-10 symptoms and the corresponding PCC values of the global PCC test are listed in table 1. This approach however has a major drawback, it assumes that the variance in the data is homogeneous across the data range.

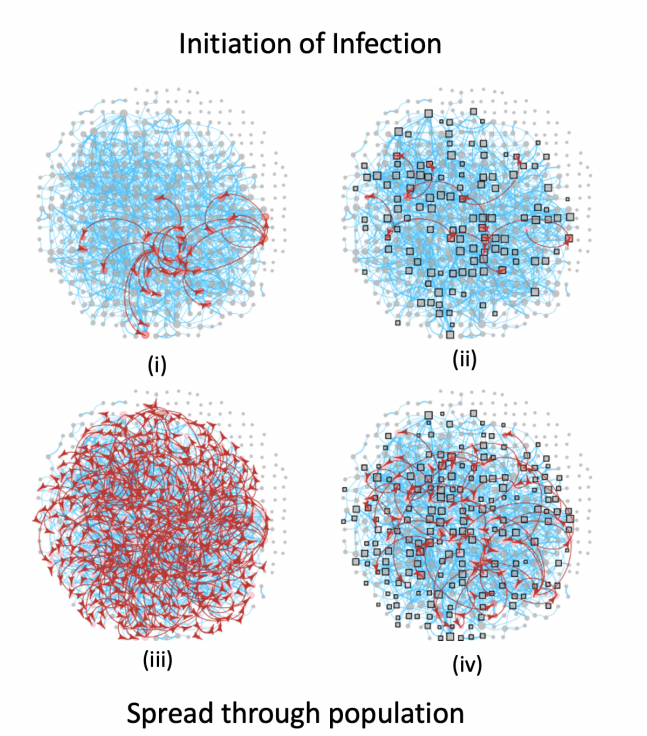

Figure 1: SIR network model , Ref: Firth et. al.

Symptom	PCC
Hypoxemia	0.643161
Eye pain	0.635909
Night sweats	0.624184
Nasal congestion	0.582434
Sinusitis	0.545885
Bradycardia	0.525026
Throat irritation	0.515732
Ageusia	0.500023
Post-nasal drip	0.491759
Anosmia	0.483690

Table 1: Global Pearson's Correlation between symptoms and covid cases

We observed that many unrelated symptoms ended up having a high correlation value. Although, these symptoms can be manually filtered, we wanted to look at how correlated the symptoms in smaller windows over the entire time frame. The table (reported in the mid-report submission) shows the symptoms which appeared the most in the top-15 most correlated symptoms. 50% indicates the median PCC value across all windows and is the metric we focus on(symptom with maximum value highlighted for each region). For the selected regions, Ageusia has the highest median PCC value majority of the times, so it is a symptom we can explore a bit more. The disease "ageusia" refers to the loss of sense of taste and is prominent

symptom among people infected with covid. Moreover, the other top symptoms are also prominent symptoms of covid. This result shows the rolling window/local PCC is an effective method to extract the most correlated symptoms.
The top 15 correlated symptoms found from the above studies are listed in table **??**

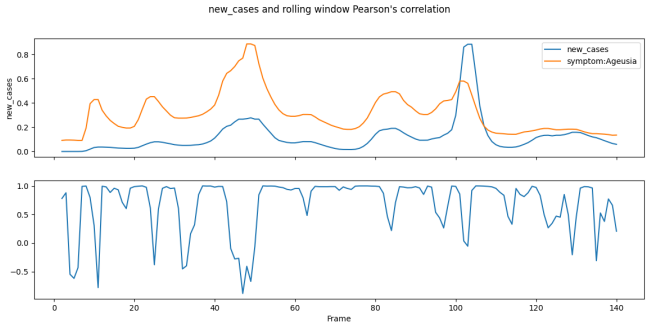

Figure 2: Plot of local window search trends of Ageusia along with the weekly covid cases for the entire nation. The bottom plot shows the PCC for the rolling window.

2. **Correlation between search trend and covid cases of neighboring states :** (Case - Georgia-Carolina, Georgia-Alabama):
   - Observation: Table 2
   - Findings:
     For a proof of concept, we performed the correlation study using google search trends from Georgia and covid cases from Alabama and South Carolina. The top 15 most correlated symptoms are listed in table **??**. Most of these symptoms also are a part of table **??** and hence shows that search trends and covid cases of neighboring states can be correlated.
3. **IMF component study:** (Case-Ageusia):
   - Observation: Figure 3
   - Findings: We performed the EMD analysis as decribed by [17], [4] and [6] to determine the statistical significance of the input data. The benefit of using the EMD method is that it can eliminate noise and non-stationary oscillations that are unrelated to features of the trend (for instance, informative searches and influential search) that do not correspond to the actual disease (Fever, Asthma etc.) or to the related disease (Covid 19). We removed the lower order IMF's and combined the first and second order IMF to form the new curve for each symptom's google trends. We again performed local PCC on the new symptom trend curves with the covid cases. We compared the correlation obtained this way with the original local PCC and observed that the correlation was comparatively lower. This suggests that the assumption we made that the oscillations were noise in the trends data was incorrect and these oscillations are important trends.
4. **Lead time study : (Case-US National, Georgia-Carolina, Georgia-Alabama):**
   - Observation: Table **??**,**??**
   - Findings: The lead time analysis was performed on the National covid (daily cases) data with respect to the symptoms

Serial	Region		
	Overall	Alabama	South Carolina
1	(Ageusia 7)	Hypoxemia	Ageusia
2	(Hypoxemia ,7)	Ageusia	Hypoxemia
3	(Low-grade fever ,6)	Low-grade fever	Anosmia
4	(Dysgeusia ,6)	Hyperthermia	Fever
5	(Fever ,6)	Fever	Eye pain
6	(Anosmia ,6)	Anosmia	Low-grade fever
7	(Chills ,6)	Headache	Headache
8	(Pneumonia ,5)	Shortness of breath	Shortness of breath
9	(Shallow breathing ,5)	Chills	Pneumonia
10	(Eye pain ,5)	Dysgeusia	Hyperthermia
11	(Shortness of breath ,4)	Common cold	Dysgeusia
12	(Headache ,4)	Eye pain	Bradycardia
13	(Shivering ,4)	Bradycardia	Infection
14	(Migraine ,3)	Weakness	Hypoxia
15	(Hyperthermia ,3)	Shivering	Migraine

Table 2: Top 15 most correlated symptoms and their corresponding occurrences and the Top 15 most correlated symptoms from Alabama and South Carolina using search trends from Georgia.

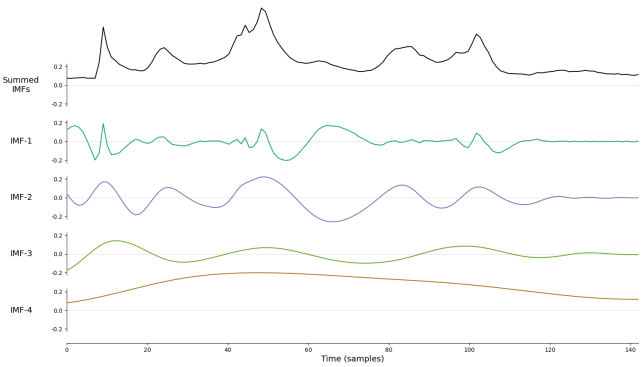

Figure 3: Google search trends for "Ageusia" divided into its IMF components

data (for year 2022, Jan to Oct; 274 dimesnion vector). We first found the most correlated symptom with the help of top-15 symptom search (using dimension reduction) and then analyse the lead time on the original data wrt. to these most correlated symptoms. Table 11 shows the result which shows logical lead time wrt. some important symptoms like 'Fever', 'Low-grade Fever' etc and many of these search trends tend not to align well with the original data (hence, we get very high lead times)
5. **Spatiotemporal Forecasting : (Case - Georgia-Carolina, Georgia-Alabama):**

Serial	Negative RTPCR		Positive RTPCR	
	Symptom	Score	Symptom	Score
1	Scoliosis	16	Common cold	36
2	Urethritis	16	Cough	36
3	Angina pectoris	15	Low-grade fever	36
4	Dysphagia	15	Acute bronchitis	35
5	Heart murmur	15	Ageusia	35
6	Hyperthyroidism	15	Anosmia	35
7	Hypomania	15	Chills	35
8	Nasal polyp	15	Hypoxemia	35
9	Neutropenia	15	Phlegm	35
10	Pelvic inflammatory disease	15	Throat irritation	35
11	Pleurisy	15	Nasal congestion	34
12	Shyness	15	Dysgeusia	33
13	Strabismus	15	Fever	33
14	Stuttering	15	Post-nasal drip	33
15	Weakness	15	Sputum	33

Table 3: Symptoms most correlated with negative and positive RTPCR tests and their correlation scores.

- Observation: Figure 4, Figure 5
- Findings:
  We initially started by using a vanilla LSTM to forecast the covid cases of a given region using the regions search trends of the most correlated symptoms extracted earlier. However, the predictions of this model were off the mark by a long distance. Hence, we started to use more advanced models using Pytorch Forecasting. Using both DeepAR and TFT, we used an encoder length of 30 to predict the next 20 timesteps. DeepAR performed well in the forecasting task but often made very inaccurate predicts but would correct them in the subsequent timesteps. This is demonstrated in figure 4. The TFT model did not have similar behavior but had similar performance and is fairly accurate (fig 5). The current work only serves as a proof of concept and we currently have not extensively tested these methods and their comparison is something which needs to be done.

6. **Rumours study about symptoms : (Case-Georgia):**
   - Observation: Table 3
   - Findings: From the above tables we can observe that Symptoms which positively correlate with the (+ve) RTPCR test reports seem to be common symptoms of covid related infections. While symptoms which corelate with (-ve) RTPCR test reports are based on fear and miss-information. Here we, see symptoms Dysphagia : Swallowing difficulties , Hypomania : Over-excitment, Heart murmur, Nasal polyp ; which are non-covid related but general health related symptoms. Thus, we find that with the healp of google search trends we can observe the general miss-information which people are having and help the general public in getting over them.
7. **Predicting vaccine requirement: (Case-Georgia):**

Symptom (a)	Score	Symptom (b)	Score	t*
Food intolerance	36	Chest Pain	37	5
Pneumonia	35	Short breath	37	5
Chills	34	Chest pain	35	5
Bronchitis	33	Globus pharyngis	35	5
Common cold	33	Middle back pain	35	5
Cough	33	Night sweats	35	5
Low-grade fever	33	Anosmia	34	5
Nasal congestion	33	Arthralgia	34	5
Acute bronchitis	32	Erectile dys.	34	5
Ageusia	32	Fever	34	5
Anosmia	32	Headache	34	5
Hemoptysis	32	Dysgeusia	33	5
Hypoxemia	32	Eye pain	33	5
Phlegm	32	Indigestion	33	5
Sinusitis	32	Migraine	33	5

Table 4: Correlation between vaccination with symptomatic search trends.

   - Observation: Tabel 4
   - Findings: From table we observe that Symptoms(a) which are directly correlated with the vaccination numbers of the state provide information about the general side-effects which people may face. These can be Chills, Food intolerance-indigestion etc. While Symtoms(b) that are top correlated symtpoms of new covid cases with search trends can provide us lead time in the vaccine requirement. This way we can better manage the pandemic.
8. **SIR-Network analysis of covid: (Case-Georgia):**
   - Observation: Figure 6.
   - Findings: The results of SIR-Network model suggest the applicability of search trends result as a proxy to actual Covid cases. In the figure we see that if we take constant parameters in the Mechanistic model then the results of the network model are not that good. But when we take time-varying parameters then we match the overall trend in Covid cases, as well as the peaks and phase transitions. And hence, the model will prove to be of great use in case where actual test data is too costly or sparsely available.

7 CONCLUSION:

We found symptoms with strong correlations between their Google search trends and actual COVID-19 cases data for US stats. We performed inter-regional studies to find We performed forecasting of neighboring state covid cases using the states google search trends(Georgia -> Alabama, Georgia -> South Carolina) and were able to accurately predict upto atleast 5 timesteps ahead. Both models DeepAR and Temporal Fusion Transformer performed well but we need to We built level contact networks using synthetic contact networks for some counties and simulated these networks using SIR model of spread with parameters estimated from GST data. This could further be expanded to other counties and simulate statewide agent-based networks using technologies such as GRADABM [3].

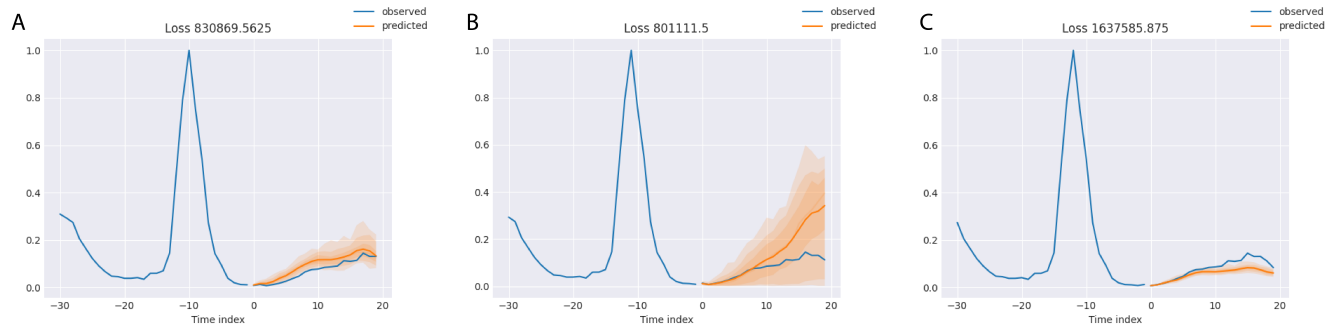

Figure 4: Forecasts for dates 2022-08-07 (A), 2022-08-14 (B), 2022-08-21 (C) for South Carolina covid cases using Georgia search trends. The DeepAR model makes a inaccurate prediction in (B) but it fixes itself in the next timestep (C).

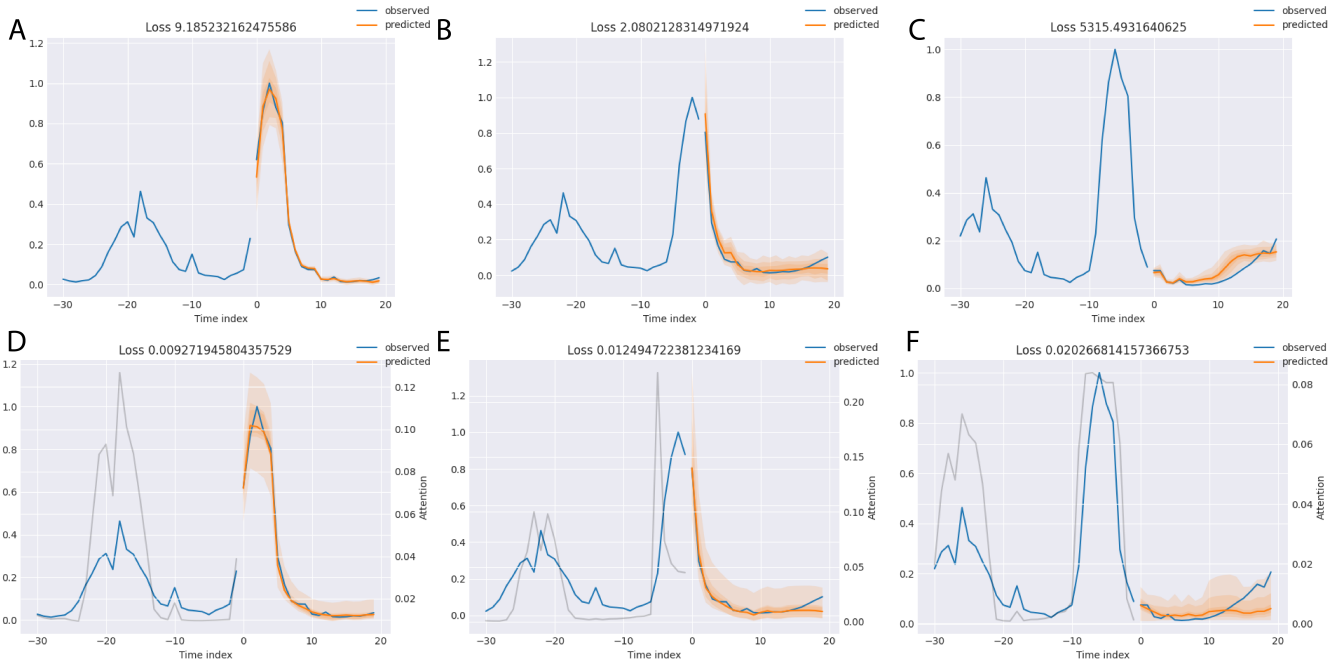

Figure 5: Forecasts for dates 2022-05-15, 2022-06-12, 2022-07-10 for Alabama covid cases using Georgia search trends. A, B & C) Forecasts of DeepAR. D, E & F) Forecasts of TFT.

REFERENCES

[1] Bijaya Adhikari, Xinfeng Xu, Naren Ramakrishnan, and B Aditya Prakash. 2019. Epideep: Exploiting embeddings for epidemic forecasting. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 577–586.

[2] Kim Asseo, Fabrizio Fierro, Yuli Slavutsky, Johannes Frasnelli, and Masha Y Niv. 2020. Tracking COVID-19 using taste and smell loss Google searches is not a reliable strategy. *Sci. Rep.* 10, 1 (Nov. 2020), 20527.

[3] Ayush Chopra, Alexander Rodríguez, Jayakumar Subramanian, Balaji Krishnamurthy, B Aditya Prakash, and Ramesh Raskar. 2022. Differentiable Agent-based Epidemiology. *arXiv preprint arXiv:2207.09714* (2022).

[4] Ramon Gomes Da Silva, Matheus Henrique Dal Molin Ribeiro, Viviana Cocco Mariani, and Leandro dos Santos Coelho. 2020. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. *Chaos, Solitons & Fractals* 139 (2020), 110027.

[5] Jacques Demongeot, Yannis Flet-Berliac, and Hervé Seligmann. 2020. Temperature decreases spread parameters of the new Covid-19 case dynamics. *Biology* 9, 5 (2020), 94.

[6] Najmul Hasan. 2020. A methodological approach for predicting COVID-19 epidemic using EEMD-ANN hybrid model. *Internet of Things* 11 (2020), 100228.

[7] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences* 454, 1971 (1998), 903–995.

[8] Atina Husnayain, Anis Fuad, and Emily Chia-Yu Su. 2020. Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan. *International Journal of Infectious Diseases* 95 (2020), 221–223.

[9] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting* 37, 4 (2021), 1748–1764.

[10] Tichakunda Mangono, Peter Smittenaar, Yael Caplan, Vincent S Huang, Staci Sutermaster, Hannah Kemp, and Sema K Sgaier. 2021. Information-seeking patterns during the COVID-19 pandemic across the United States: Longitudinal analysis of Google Trends data. *J. Med. Internet Res.* 23, 5 (May 2021), e22933.

Application	Data type	Publisher	Link
COVID-19 cases data	COVID-19 statistics	Johns Hopkins University	github/CSSEGISandData/COVID-19
COVID-19 cases data	COVID-19 statistics	CDC	cdc/2019-ncov
Measuring emotions	Textual/Embeded data	COVID-19 Real World Worry	github/ben-aaron188/covid19worry
Media Coverage	National and International news article and interpretation	Humanities and social sciences communication [nature]	nature/News-media-coverage-of-COVID-19
Search trends data	Normalised keyword search data	Google	google/search-trends
COVID-19 Trends and Impact Survey	Symptoms and behaviour data	Facebook and Delphi CMU	cmu-delphi/symptom-survey
Hospital Capacity Data	Number of beds occupied and available state wise data	U.S. Department of Health & Human Services	healthdata/capacity-bed-data
Contact network	county-level directional synthetic contacts	Biocomplexity Institute, University of Virginia	dataverse.lib.virginia.edu
COVID-19 vaccination data	COVID-19 statistics	USCDCP	ourworldindata

Table 5: Data Sources

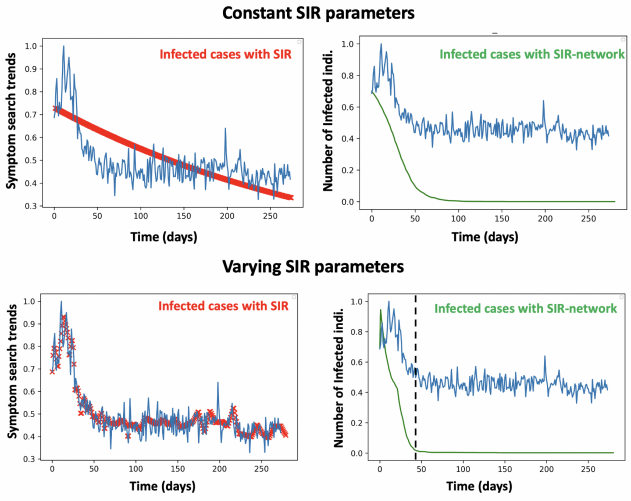

Figure 6: Comparison of covid cases and SIR-network model results.

National data	Top k symptoms	Match	Lead time (t*)
1	Anosmia	18	271
2	Acute bronchitis	17	267
3	Ageusia	17	271
4	Bronchitis	17	267
5	Chills	17	271
6	Common cold	17	267
7	Cough	17	267
8	Dysgeusia	17	270
9	Fever	17	14
10	Hemoptysis	17	270
11	Hypoxemia	17	271
12	Low-grade fever	17	7
13	Nasal congestion	17	271
14	Pneumonia	17	6
15	Post-nasal drip	17	15

Table 6: Lead Time of symptom search trends on national weekly covid cases.

[11] Amaryllis Mavragani et al. 2020. Tracking COVID-19 in Europe: infodemiology approach. *JMIR public health and surveillance* 6, 2 (2020), e18941.
[12] Amaryllis Mavragani and Konstantinos Gkillas. 2020. COVID-19 predictability in the United States using Google Trends time series. *Scientific reports* 10, 1 (2020), 1–12.
[13] Network Systems Science And Advanced Computing Division. 2020. Population contact rates by age and county. https://doi.org/10.18130/V3/OMEPOT
[14] Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Waqar Aslam, and Gyu Sang Choi. 2020. COVID-19 future forecasting using supervised machine learning models. *IEEE access* 8 (2020), 101489–101499.
[15] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2020. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* 36, 3 (2020), 1181–1191.
[16] Stefan Thurner, Peter Klimek, and Rudolf Hanel. 2020. A network-based explanation of why most COVID-19 infection curves are linear. *Proc. Natl. Acad. Sci. U. S. A.* 117, 37 (Sept. 2020), 22684–22689.
[17] Albert C Yang, Jong-Ling Fuh, Norden E Huang, Ben-Chang Shia, Chung-Kang Peng, and Shuu-Jiun Wang. 2011. Temporal associations between weather and headache: analysis by empirical mode decomposition. *PloS one* 6, 1 (2011), e14612.
[18] Abdelhafid Zeroual, Fouzi Harrou, Abdelkader Dairi, and Ying Sun. 2020. Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos, Solitons & Fractals* 140 (2020), 110121.