Removing Media Coverage Bias from COVID-19 Symptom Search Trends Data

Siddhartha Vemuri
svemuri8@gatech.edu

Shen En Chen
achen353@gatech.edu

Mohit Aggarwal
mohit7@gatech.edu

ABSTRACT

Our project plans to analyze the impact of media coverage on the predictive power of Google Search Trends Relative Search Volume (RSV) of different COVID-19 symptoms for reported positive cases in the State of New York. We will first perform analysis of the correlation/similarity between the RSV of several COVID-19 symptoms and the reported COVID-19 cases for the State of New York using several metrics including Spearman's Rank Correlation, Dynamical Correlation, and the Dynamic Time Warping Algorithm (DTW). Then, we will evaluate the use of a Recurrent Neural Network (RNN) as a method to remove the noise introduced by the signal of media coverage from RSV and see if it is better able to correlate with true COVID-19 case counts. We will compare this with a baseline method of autoregression as implemented by Lampos et al. [10] and simple linear regression. The effectiveness of these methods will be evaluated using (1) Pearson Product-Moment Correlation (2) Spearman Rank Correlation, and (3) the Dynamic Time Warping Distance.

1 INTRODUCTION

Google Search Trends has been used extensively by researchers in the past as an epidemiological tool to study the spread of disease. From gauging the health impact of heat waves [9], forecasting influenza-like illness (ILI) [18], to predicting COVID-19 cases [2]. In particular, various studies have been published that analyze the correlations between Google Relative Search Volume (RSV) of COVID-19 symptoms and confirmed COVID-19 cases, as well as the predictive value of the former for the spread of COVID-19. RSV is a metric that represents the search volume of a keyword as a function of how often it appears relative to the total search volume for the day (the exact calculation is not made public). However, studies such as [15] showed that these GT-based literatures often overlook the confounding effects of COVID media coverage on symptom RSV, leading to potential false-positive results. For this project, we aim to (1) analyze the correlation between COVID media coverage and symptom RSV for common (e.g., cough, fever, etc.) and COVID-19 specific symptoms (e.g., anosmia, shortness of breath, etc.), (2) statistically denoise or decorrelate the confounding effect of media coverage on symptom RSV, and (3) evaluate the effectiveness of the proposed denoising method.

2 RESPONSE TO MILESTONE COMMENTS

As seen in 1, there is an anomalous spike in media coverage of COVID symptoms (as shown by the aggregated symptom RSV signal which will be further described later in the report) around mid-April that seemingly causes RSV to increase relative to the actual reported COVID cases (which in fact decreases). In this case, we see how RSV may not be indicative of the state of disease spread, but rather influenced by the confounding effects of media coverage.

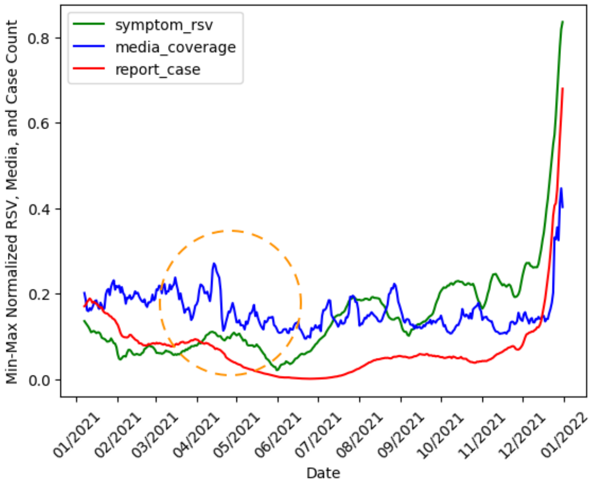

Figure 1: Example of RSV anomaly seemingly caused by media bias

These sorts of anomalies reduce the reliability of Google Search Trends as an epidemiological indicator, which makes it all the more important to find an approach to overcome them.

3 RELATED WORK

COVID-19 is one of the largest pandemics in human history and the deadliest in the digital age. With increased access of data through the internet and the release of the Google COVID-19 Search Trends symptoms dataset [11], researchers and government authorities have extensively studied the feasibility of using Google Trends as an infoveillance tool.

3.1 Google Trends as COVID-19 Predictor

The most common quantitative analysis on Google Trends for COVID-19 is to examine its predictive power for COVID-19 cases. [2] examined the relationships between the top nine Google search trends for COVID-19 symptoms and confirmed and death cases of COVID-19 during the first year of the pandemic in all states of the United States through dynamic correlation. The results reveal the efficacy of Google search queries to forecast COVID-19 cases and mortality for up to three weeks in advance. However, the authors only analyzed the COVID-19 confirmed and death cases from the start of the pandemic until the end of October 2020, and part of their analyses identified different trajectory patterns of the cases over time, signaling that the patterns might change should the study time interval be extended. Furthermore, because the study was carried out during the first wave of the pandemic, COVID-19 cases were likely to be underestimated due to insufficient testing in most parts of the US.

Contrary to the work discussed above, [14] showed that the daily relative search volumes (RSV) of COVID-related Google searches are an unreliable predictor for COVID-19 cases. The authors surveyed Google search trends in (1) Italian regions and cities and (2) countries and cities worldwide during two overlapping periods of time in 2020 and analyzed the Pearson and Spearman correlations between Google Trends daily RSVs and COVID-19 cases. For both Italian regions and countries around the world, the correlation underwent significant variations between the two time spans of interest, signifying the variability of the correlation between the two quantities. Moreover, the authors observed a large of amount of anomalies (missing values) in the GT data that could invalidate the statistical significance of the correlation.

3.2 Impact of Media Coverage on COVID-19 Search RSV

There may be many confounding factors that impact the correlation between RSV and COVID-19 cases. A major example of such is media coverage: a sudden increase of media coverage of COVID-19 symptoms, particularly those that are rare for common influenza, can potentially lead to a rise in RSV as the public seek to self-diagnose themselves or take precautions. [15] examined the Granger causality of Google Trends RSVs to weekly COVID-19 positivity in eight different English-speaking countries and Japan and analyzed the impact of media coverage on the causality mentioned above. While the authors identified “sense of smell” and “loss of smell” as the most reliable GT keywords across all the evaluated countries, they also found the search trends to be more aligned with media coverage than reported COVID-19 cases. The Granger causality of these two keywords in COVID-19 cases is weakened after adjusting search trends with media coverage, suggesting that search trends may be much more strongly correlated with media coverage than COVID-19 positivity trends. Similar results were reported by [4]. The authors found that world-wide searches of most of the COVID-19 symptoms such as shortness of breath, headache, chest pain, and sneezing had strong correlations with the daily confirmed cases and deaths from COVID-19 with clear negative lags but searches for anosmia and dysgeusia peaked with positive lags and strongly correlate with the announcement of the loss of smell as a potential marker for the COVID-19 infection by the international medical community on March 20, 2020.

To narrow down the focus, [16] assessed whether the searches for ageusia and anosmia, two COVID-specific symptoms that had not come into public attention until the pandemic, were primarily correlated to media coverage or COVID-19 positivity trends. The authors observed synchronized rises and peaks of search trends across 17 different countries irrespective of their epidemic situations and how they coincided with the media announcement of these COVID-19 symptoms. Prior to the announcement, GT data reflected COVID-19 cases and deaths with variable but reliable correlations.

Before the pandemic, [4] compared the reliability of GT of common diseases with lower media coverage, and for less common diseases attracting major media coverage in Italy. The results indicated that GT has moderate reliability in monitoring and predicting epidemiological trends in common diseases with poor media coverage. In general, GT is more influenced by media clamor, enough to call the reliability of epidemiological predictions into question.

3.3 Mitigating Media Effect on Symptom RSV

Although many have identified media coverage as a major influencing factor of online search trends for diseases and symptoms, few have evaluated the feasibility of denoising GT data with media coverage over time and using it as a more effective indicator of epidemiological trends. [10] is one of the few literatures that mitigated the media effect when using online search trends to track COVID-19. The authors leveraged two autoregressive models—one forecasts the search trend signal of the current day t, g_t, using its previous values g_{t-1} and g_{t-2} with an absolute error ϵ_t^1 and the other forecasts g_t using the media signals of t and the previous two days, m_t, m_{t-1} and m_{t-2} in additional to g_{t-1} and g_{t-2} with an absolute error ϵ_t^2. The effects of media is quantified day by day through $\gamma = \frac{\epsilon_t^2}{\epsilon_t^1}$ and search trend is denoised by reweighting the search trend signal with the γ of the day. However, as this article does not focus primarily on removing the effect of media coverage from RSV data, this work seems to be somewhat limited in its evaluation of how well the approach performs in removing the effects of media bias from RSVs. Thus, our work will additionally provide a quantitative analysis of Lampos et al.'s methods and use them as a baseline by which to measure the performance of our experimental RNN-based method.

4 MAIN OBJECTIVE

Our main objective is to evaluate the use of Linear Regression and Recurrent Neural Networks to remove the impact of media coverage on Google RSV for COVID-19 symptoms as a predictive signal for true COVID case counts against a baseline method of autocorrelation as described by Lampos et al. [10]. We aim to conduct this evaluation by first finding the correlation between RSV and COVID-19 case count signals, then the correlation between the processed RSV signals according to Lampos et al.'s methods and COVID-19 case counts, then the predictive power of simple linear regression using RSV and media coverage to predict case counts, and finally, the correlation between the RSV and media signals processed by an RNN and the COVID-19 case counts.

5 APPROACH

5.1 Data and Data Collection

Our project includes three major sources of data: the Google COVID-19 Search Trends symptoms dataset that contains RSV for several COVID-19 symptoms [11], data taken from the MediaCloud API which allows us to find the number of news articles written about a particular keyword each day, and reported daily positive cases in New York State taken from NY State's website.

For all data sources, We consider data from January 1, 2021 to December 31, 2021. This is because case reporting was very inaccurate for large portions of 2020 due to the shortage of reliable testing. 2021 has relatively reliable COVID-19 case count reporting which allows us to more effectively evaluate the correlation between true COVID-19 cases and symptom RSV without having to consider sudden increases in the effectiveness of testing making the case count dataset more aligned with true case counts.

For data collection, the search trends are publicly available online and we focus on New York State daily search trends for 14 different symptoms listed by the CDC as the symptoms of COVID-19 [3]. These include: cough, fever, ageusia (lost of smell), anosmia (loss of taste), sore throat, chills, diarrhea, fatigue, headache, nausea, vomiting, shortness of breath, muscle weakness, and congestion.

For news articles, we consider the daily counts of media articles published that contain the corresponding symptoms as keywords in either the title or article body. We collected the data using the Media Cloud database [7], which is a free initiative that aggregates media articles from a wide variety of platforms and provides more flexibility on query specification than `pygooglenews`. We set the region to the US and queried nationwide relevant new articles as a proxy of the media coverage for New York State residents.

For the daily COVID-19 case counts of New York State, we fetched the data from New York State Statewide COVID-19 Testing API [13]. As discussed prior, we considered only COVID-19 case counts rather than hospitalization or death count data due to treatments and vaccines that rolled out during the period that would greatly change the fundamental nature of the signals while we evaluate their correlation with RSV.

5.2 Evaluation Metrics

A major point of consideration for our work is how we will evaluate the "goodness" of any of our approaches. As we are facing the non-trivial problem of attempting to ensure that there is a proper correspondence between trends in RSV data and trends in case counts, we must make an appropriate choice of the metrics we use to quantitatively measure this. Thus, we consider the following metrics to evaluate the similarity between the studied signals:

5.2.1 Root-Mean-Square Error. The root-mean-square error is the square root of the average of squared errors lower-bounded by 0. As this is a well-known metric with a self-explanatory definition, we omit the mathematical explanation here. We provide this metric for the sake of measuring model performance in a more traditional, accuracy-based manner, but we do not consider it in-depth for our analysis as it is not robust to temporal shifts between signals.

5.2.2 Pearson Correlation. The Pearson Correlation [5] measures how two continuous signals co-vary over time and indicate the linear relationship as a number between the range [-1, 1]. Since we have visited this concept many times in the assignments of class, we also omit the mathematical definition of the coefficient.

5.2.3 Spearman Rank Correlation. Formally, the Spearman Rank Correlation is denoted by r_s and is a numerical value between [-1,1]. The role of r_s is to measure the likelihood of a variable increasing as the other increases which is called direct association. And same is the case when one variable decreases it measures the likelihood of the other decreasing which is called inverse association. A direct association corresponds to positive values and a negative value corresponds to a negative association. Furthermore, a value of 0 just indicates there is no relation. The formula mentioned below is used to calculate Spearman's Rank Correlation:

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the two ranks of each observation and n is the total number of observations.

Ageusia	Anosmia	Chills
30.303	13.378	13.715
Cough	Diarrhea	Fatigue
10.272	56.459	88.217
Fever	Headache	Muscle Weakness
9.565	112.916	159.547
Nasal Congestion	Nausea	Shortness of Breath
12.624	81.959	137.923
Sore Throat	Vomiting	Aggregated
12.070	62.518	10.29

Table 1: Minimum sum of distances for Dynamic Time Warping

5.2.4 Dynamic Time Warping (DTW). The Dynamic Time Warping (DTW) algorithm [1] can be used to measure the similarity between two curves while accounting for temporal shifts between them, making it especially useful when working with leading or lagging signals.

[5]. Formally, given two time series i and j with the subscripts representing the timestep, we want to minimize the distance. Here d represents the euclidean distance. At the same time we use the values which were computed from the previous D_{min} to obtain the value for the subsequent points. The warping path using this formula is found using a dynamic programming approach and there are various constraints like boundary, monotonicity applied to make it computationally less expensive.

$$D_{min}(i_k, j_k) = \min_{i_{k-1}, j_{k-1}} D_{min}(i_{k-1}, j_{k-1}) + d(i_k, j_k | i_{k-1}, j_{k-1})$$

Our approach will use the distance value calculated by applying DTW to the outputs of our models and the true positive COVID case counts. The lower the distance, the more similar the outputs of our models and true case counts are, giving us insights into how effective our models are at accomplishing their objective of denoising the RSV signal to better correlate with the case count signal.

5.3 Preliminary Analysis

Our preliminary analysis will focus on finding the extent to which RSV signals are appropriate predictors of COVID-19 case counts without removing the confounding effect of media coverage.

Table 1 presents the minimum sum of DTW distances required to align the RSV with reported cases for each symptom after min-max scaling both the RSV data and case count data. The symptoms with the top-4 lowest (best) scores are fever, cough, sore throat, and nasal congestion (shown in Figure 2), all of which are common symptoms

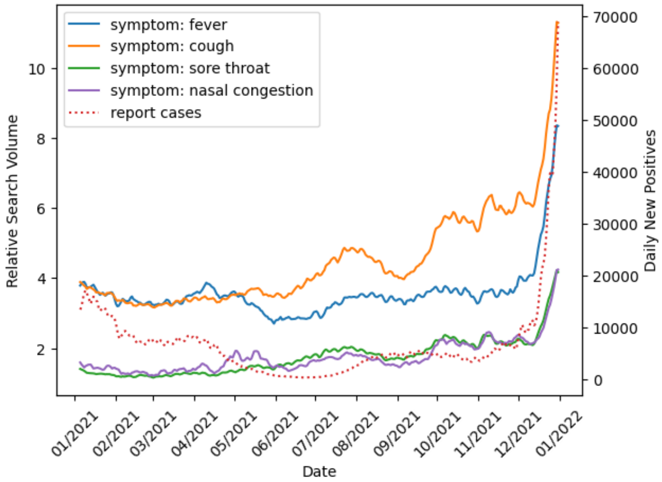

Figure 2: Symptoms with top-4 similar RSV to reported cases

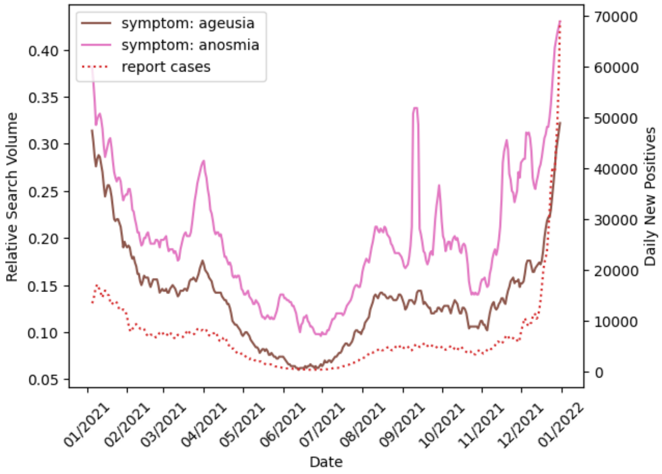

Figure 3: RSVs of ageusia and anosmia compared to reported cases

for ILI and common cold. RSVs of more COVID-specific symptoms such as anosmia (loss of tatse) and ageusia (loss of smell) show high to medium similarity with the reported case curve (Figure 3). Specifically, symptoms with medium similarity often contain local peaks around April and September when the reported cases also increase mildly. We believe that these are cases when confounding factors such as media coverage caused the RSV to overshoot and erroneously indicate a change in disease trend. This demonstrates that there is sufficient motivation for our work which will directly attempt to remove the confounding effects of media coverage on RSV data.

5.4 Aggregating RSVs by Symptom Frequency

To effectively leverage all individual symptom RSVs, we combine them into one single time series by taking the weighted sum of the RSV value at each time step. We use the reported symptom frequencies from the CDC Morbidity and Mortality Report released in 2020 [8] as listed in 2 as the weights for different symptoms. This method also allows us better align with the methodology of previous work [10].

In the following sections, we will examine different approaches to remove the confounding effects of media bias from the aggregated RSV signal.

Ageusia	Anosmia	Chills
0.20	0.20	0.63
Cough	Diarrhea	Fatigue
0.84	0.38	0.62
Fever	Headache	Muscle Weakness
0.80	0.59	0.63
Nasal Congestion	Nausea	Shortness of Breath
0.10	0.34	0.57
Sore Throat	Vomiting	
0.40	0.13	

Table 2: Reported frequencies of symptoms in study of COVID patients

5.5 Baseline: Lampos et al.

One of the baselines of our approach is the method proposed by [10]. Essentially, the method denoises the aggregated RSV time step by time step by constructing two linear regression models that predict the current-day aggregated RSV with 2 lags of past values with and without the addition of media trends. We defer the details of this method to 3.3. We re-implemented the model based on the description and mathematical formulation in the publication. We first analyze the default model in the original publication with 2 lags and analyze the optimal number of lags for different evaluation metrics.

5.6 Linear Regression

Another approach we took was the application of a linear regression model. We test and evaluate four versions of a linear regression model using lead times ranging from 1-30 days that attempt to predict case counts. These four versions of the model are as follows considering k as the lead time:

(1) Given RSV data for day *i*, predicts case counts on day i+k
(2) Given RSV and case count data for day *i*, predicts case counts on day i+k
(3) Given RSV and Media data for day *i*, predicts case counts on day i+k
(4) Given RSV, Media, and case count data for day *i*, predicts case counts on day i+k

The goal of this approach was to evaluate whether the simple weight assigned by a linear regression model to a predictor variable representing the media bias would be sufficient to remove its confounding influence from the RSV data and allow it to more accurately predict case counts. Both choices of whether or not to include case count data from the same day RSV and media data came from were explored for the sake of completeness.

We believed that this naive approach would have limited performance in denoising because it ignores the sequential and temporal relations between consecutive time steps. However, since we found no existing literature adopting this approach, we believed it would be worthwhile to explore and compare.

5.7 Recurrent Neural Network (RNN)

The main approach we explore is a recurrent neural network (RNN). RNNs have been used to generate incredibly precise predictions on sequential data. They are ideal for using time-series data, such as the data we need to examine since they can capture temporal relationships and variable-length observations [17].

Our specific architecture utilizes a simple RNN with 8 hidden units and ReLU activation. It has a total of 96 learnable parameters, which does not allow for much complexity, but seems to be suitable for the small dataset size (<300 observations). We train and evaluate several RNN models using this architecture for a variety of sequence lengths (ranging from 2-14), with and without including media data as an input feature. The goal of this is to identify the improvement in metric scores that including media data as an input along with RSV data is able to provide.

Our expectation is that model should learn some function that weights the contributions of media coverage negatively while weighing the contributions of RSV positively in order to more accurately predict the COVID-19 case count label at each step. We hope this will effectively remove the confounding effect of media coverage on RSV.

5.8 Motivating Questions

- Is the baseline Lampose et al. approach sufficient for denoising?
- Is the single beta parameter weight learned by linear regression enough to negatively weight and effectively denoise the media signal from the RSV signal?
- For linear regression, can we yield better performance if we consider lead time, allowing the model to learn to predict $k - th$ day into the future instead of the immediate next day?
- Would learning the sequential/auto-regressive relationships between the RSV, media coverage, and reported case counts provide useful information for the model to forecast case counts, so as to achieve denoising?
- Is there sufficient training data for the RNN to learn a robust function that doesn't overfit or underfit?
- Is the RNN architecture too simple to learn anything meaningful about the relationship between media coverage, RSV, and case counts?
- How will each of the studied methods perform on our suite of evaluation metrics and what insights can we extract from their individual performances and performances relative to each other?
- After comparing all the methods, are the metrics we defined sufficient for the articulated objective–measuring the effectiveness of denoising?

6 RESULTS

In this section, we present the results of each of the models described in Sections 5.5 - 5.7 as well as the answers to the motivating questions described in Section 5.8 for our experiments.

6.1 Baseline: Lampos et al.

The performance of Lampos et al. is reported in Table 3. We extended the model by allowing a variable number of k lag days. The lag day k means the number of days of past RSV and media coverage we use to denoise the reported case counts of the current day. We ran the model for $k = 1$ to $k = 30$ and identified the best performances for different metrics and their corresponding k's (bolded). We also include performances for $k = 2$, the original configuration in [10].

Lag	RMSE	Pearson	Spearman	DTW
0	11894.05	0.75 ($2.58 \cdot 10^{-69}$)	0.19 ($1.20 \cdot 10^{-4}$)	10.29
1	11884.87	**0.52** ($2.24 \cdot 10^{-26}$)	0.18 ($3.42 \cdot 10^{-4}$)	125.73
2	11886.48	0.49 ($4.12 \cdot 10^{-13}$)	0.17 ($1.43 \cdot 10^{-3}$)	111.36
5	11870.53	0.43 ($2.14 \cdot 10^{-17}$)	**0.19** ($2.42 \cdot 10^{-4}$)	116.94
29	**11668.07**	0.29 ($2.96 \cdot 10^{-8}$)	0.14 ($8.44 \cdot 10^{-3}$)	**84.57**

Table 3: Performance of Lampos et al. [10] with different lags. Lag = 0 days represents the performance without applying denoising. For Pearson and Spearman rank correlations, we format the results as *score (p-value)*.

For $k = 1$, $k = 2$, and $k = 5$ lag days, the Pearson correlation drops with increased p-values, indicating a reduction in correlation between the RSV and the reported case counts. For Spearman rank correlation, the correlation fluctuates within the same magnitude. In terms of the distance for DTW, we see a drastic increase compared to the baseline. For $k = 29$, we report the results because it yields the lowest RMSE and DTW distance among all variations of Lampos's method. However, due to the large value of k, we believe the results provide few practical values.

To visualize the denoised RSV, we normalized each resulting time series as well as the reported case counts independently between 0 and 1 and project them onto the same plot, as shown in Figure 4. Qualitatively, after applying Lampos's method to denoise the RSV, the time series becomes more compact vertically and the spike toward the end of the year becomes less observable. This signals that the RSV becomes less indicative of a potential spike and its magnitude after denoising and explains the drop in correlation and increase in DTW distance.

6.2 Linear Regression

We report the performance of linear regression in Table 4. Our findings show that k=1 day of lead time (using data from a day to predict the case counts of the next day) gives us the best values across the different metrics we evaluated. However, we elected to include results for a lead time of $k = 7$ days as well as we feel it

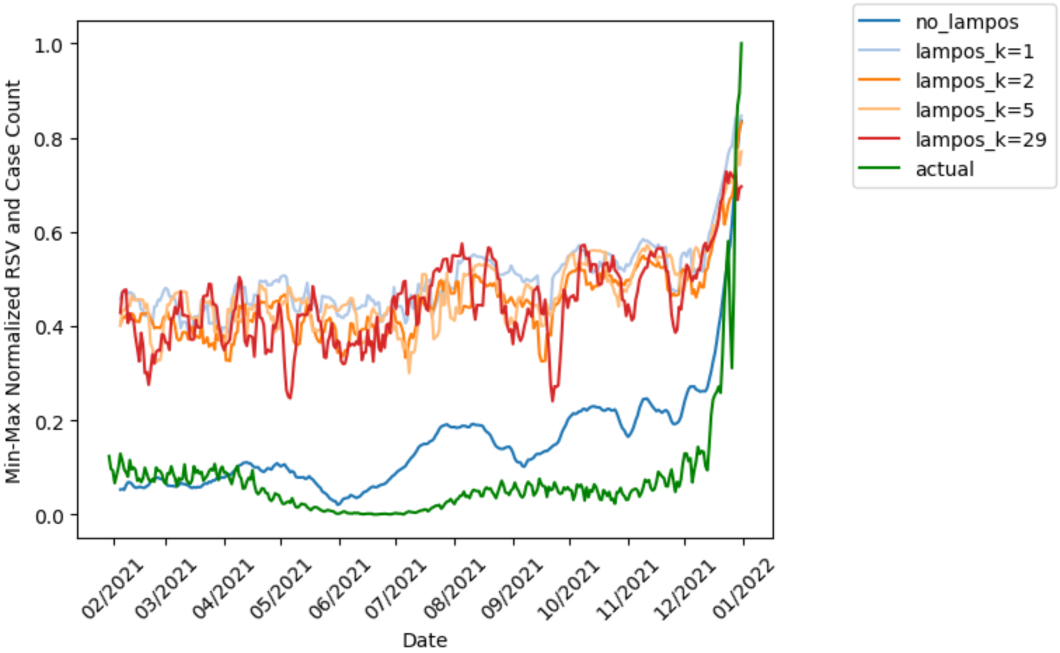

Figure 4: Normalized denoised RSV and reported case counts. For all trials of Lampos experiments, we min-max normlize the denoised RSV as well as the reported case counts and plot them together. The values are smoothed with a 7-day rolling window after normalization.

reflects a good balance of performing well on our metrics while being useful from a policy standpoint to predict disease trends.

As we can see, the linear regression model does not experience a significant improvement in performance when including media data. DTW distance, Pearson score, and Spearman score are all within .01 of each other, and RMSE is not significantly different between the two. This seems to indicate that the linear regression model assigning a singular weight to media data is insufficient to appropriately denoise its influence from the RSV data. This supports our intuition that a more complex method is needed to negatively weight media bias in order to properly denoise the media signal from the RSV signal.

An interesting point to note is that providing the linear regression model with the case count data from the same day the RSV and media data come from improves its performance on all metrics collected. This is likely due to the fact that the model is able to incorporate information about a good "baseline" value to start the prediction from and adjust according to the values of the other predictor variables. This essentially adds an autoregressive component to the linear regression model which may be worth exploring in future work.

6.3 Recurrent Neural Network (RNN)

The results of using RNNs to predict case count are reported in Table 5. As the sequence length in the training data increases, the RMSE generally increases when disregarding whether the media data is included; one potential reason is that including too many past values could introduce more noise to the RNN to learn sequential relationships of the aggregated RSV, media data, and the reported cases. For the Pearson correlation, we can see that the statistic is relatively stable across different trials. But the Spearman rank correlation is only stable when media data is included; without media data, the correlation drops from 0.93 with a sequence length of 2 to 0.83 for a sequence length of 0.83.

When including media data, the Pearson correlation does not change much, but the Spearman rank correlation increases for most of the sequence lengths, and the minimum distances for DTW decrease more significantly than the decrease observed on linear regression. Comparing the results to that Table 4, we see that the minimum distance for DTW using an RNN with media data included is much lower than that for linear regression, showing that the increased model complexity and capability of RNN in learning sequential relationships was able to better leverage the media data to predict case counts that are more similar to the ground truth in trend.

7 FUTURE WORK

For future work, we plan to delve further into other metrics used in related work and explore more on using neural networks to denoise the RSV.

7.1 Other Metrics

7.1.1 Cross Correlation. When researching approaches to calculate lead time, we encountered cross-correlation as a common method to measure the time shift in two signals. We experimented with performing cross-correlation between COVID-19 case counts and

Lead Time	With Media Data	With Current Case Count	RMSE	Pearson	Spearman	DTW
1	Yes	Yes	2359.95	0.98 (1.73 $\cdot 10^{-119}$)	0.91 (2.02 $\cdot 10^{-69}$)	2.31
1	No	Yes	2674.86	0.97 (2.75 $\cdot 10^{-108}$)	0.92 (7.53 $\cdot 10^{-74}$)	2.32
1	Yes	No	3153.83	0.98 (3.29 $\cdot 10^{-119}$)	0.86 (3.40 $\cdot 10^{-53}$)	4.07
1	No	No	3353.88	0.97 (5.95 $\cdot 10^{-110}$)	0.87 (9.89 $\cdot 10^{-58}$)	4.16
7	Yes	Yes	3575.71	0.95 (9.53 $\cdot 10^{-92}$)	0.86 (3.81 $\cdot 10^{-53}$)	3.20
7	No	Yes	3554.38	0.95 (6.58 $\cdot 10^{-92}$)	0.87 (9.69 $\cdot 10^{-56}$)	3.62
7	Yes	No	4675.03	0.94 (4.49 $\cdot 10^{-84}$)	0.81 (1.68 $\cdot 10^{-41}$)	4.35
7	No	No	4684.23	0.94 (4.23 $\cdot 10^{-83}$)	0.84 (6.39 $\cdot 10^{-47}$)	4.25

Table 4: Performance of linear regression for lead times of 1 and 7, including and excluding media data, and including and excluding current day case count data (all combinations tested). When looking at the Pearson and Spearman columns, note that values are formatted as *score (p-value)*.

Sequence Length	With Media Data	RMSE	Pearson	Spearman	DTW
2	Yes	5185.83	0.96 (3.09 $\cdot 10^{-43}$)	0.94 (2.07 $\cdot 10^{-34}$)	0.90
2	No	6322.54	0.95 (3.47 $\cdot 10^{-37}$)	0.93 (2.77 $\cdot 10^{-32}$)	0.97
4	Yes	5468.47	0.97 (7.42 $\cdot 10^{-44}$)	0.94 (7.71 $\cdot 10^{-34}$)	1.04
4	No	4937.88	0.97 (1.29 $\cdot 10^{-43}$)	0.94 (1.03 $\cdot 10^{-33}$)	0.88
5	Yes	6501.54	0.97 (6.94 $\cdot 10^{-44}$)	0.94 (4.93 $\cdot 10^{-35}$)	1.37
5	No	6143.85	0.97 (1.09 $\cdot 10^{-43}$)	0.93 (7.23 $\cdot 10^{-32}$)	1.18
7	Yes	5580.93	0.96 (3.89 $\cdot 10^{-39}$)	0.93 (7.42 $\cdot 10^{-30}$)	0.87
7	No	11861.85	0.95 (4.15 $\cdot 10^{-34}$)	0.90 (1.05 $\cdot 10^{-23}$)	1.85
10	Yes	7416.88	0.97 (1.11 $\cdot 10^{-39}$)	0.95 (6.81 $\cdot 10^{-33}$)	1.00
10	No	16315.65	0.87 (1.12 $\cdot 10^{-20}$)	0.79 (5.55 $\cdot 10^{-15}$)	2.75
14	Yes	7042.55	0.95 (7.25 $\cdot 10^{-32}$)	0.94 (1.12 $\cdot 10^{-28}$)	0.81
14	No	15608.15	0.77 (4.85 $\cdot 10^{-13}$)	0.83 (2.38 $\cdot 10^{-16}$)	2.03

Table 5: Performance of the recurrent neural network. We format the Pearson and Spearman columns as *score (p-value)*.

well-correlated symptom RSVs but were unable to obtain meaningful results despite our best efforts. We would like to continue looking into cross-correlation as we believe it could potentially be a powerful way to quantify whether a denoised version of RSV gives us more advance notice on trends in true COVID-19 case counts.

7.1.2 Dynamic Conditional Correlation (DCC). Literature such as [6] analyzed the time-varying correlation of time series using advanced dynamic conditional correlation (DCC) in additional to static measures such as Pearson correlation and Spearman Rank Correlation. The specific model that we found to achieve this is Generalized AutoRegressive Conditional Heteroskedasticity (GARCH). It is often used in analyzing time series data such as stock market volatility where the variance error is believed to be serially autocorrelated. While it is unclear whether this is the approach [6] used, the authors did show that advanced DCC models provides better insights than rolling window correlation. We attempted to calculate this with existing implementation [12] but were not able to obtain accurate results.

7.2 Other Approaches

As future work, we believe that more can be explored for RNNs. During the experiments, we noticed the training is rather unstable, possibly due to (1) the insufficient number of features in the input sequences and (2) the lack of regularization to prevent over-fitting. Another potential future work is to extend this work on denoising RSV for more influenza-like illnesses (ILI), which would allow us to obtain a larger volume of data over years.

8 TEAM MEMBER CONTRIBUTIONS

- **Sid**
  - Collected and processed case count data
  - Conducted preliminary lead time analysis experiments
  - Created Aggregated Signals
  - Conducted Linear Regression experiments
  - Created Plotting module
- **Andrew**
  - Collected and processed RSV data
  - Processed media coverage data
  - Conducted preliminary correlation analysis experiments
  - Conducted Baseline: Lampos experiments
  - Created Evaluation module
- **Mohit**
  - Collected media coverage data
  - Analyzed the differences and biases between different search keywords including different phrasing of the same symptoms and different geographical conditioning
  - Conducted the RNN experiment, including:
    - Studied forecasting time series with RNN
    - Explored different depths of RNN and LSTM architectures
    - Fine-tuned the hyper-parameters for training on different architectures

REFERENCES

[1] 2007. *Dynamic Time Warping*. Springer Berlin Heidelberg, Berlin, Heidelberg, 69–84. https://doi.org/10.1007/978-3-540-74048-3_4

[2] M. Abbas, T. B. Morland, E. S. Hall, and Y. El-Manzalawy. 2021. Associations between Google Search Trends for Symptoms and COVID-19 Confirmed and Death Cases in the United States. *Int J Environ Res Public Health* 18, 9 (04 2021).

[3] cdc. 2022. *Symptoms of COVID-19*.

[4] G. Cervellin, I. Comelli, and G. Lippi. 2017. Is Google Trends a reliable tool for digital epidemiology? Insights from different clinical settings. *J Epidemiol Glob Health* 7, 3 (09 2017), 185–189.

[5] Jin Cheong. 2019. *Four ways to quantify synchrony between time series data*.

[6] H. Cinarka, M. A. Uysal, A. Cifter, E. Y. Niksarlioglu, and A. Çarkoğlu. 2021. The relationship between Google search interest for pulmonary symptoms and COVID-19 cases using dynamic conditional correlation analysis. *Sci Rep* 11, 1 (07 2021), 14387.

[7] Media Cloud. 2022. *Media Cloud is an open-source platform for media analysis.*

[8] Burke et al. 2020. Symptom Profiles of a Convenience Sample of Patients with COVID-19 — United States, January–April 2020. *Morbidity and Mortality Weekly Report* (2020).

[9] Helen K. Green, Obaghe Edeghere, Alex J. Elliot, Ingemar J. Cox, Roger Morbey, Richard Pebody, Angie Bone, Rachel A. McKendry, and Gillian E. Smith. 2018. Google search patterns monitoring the daily health impact of heatwaves in England: How do the findings compare to established syndromic surveillance systems from 2013 to 2017? *Environmental Research* 166 (2018), 707–712. https://doi.org/10.1016/j.envres.2018.04.002

[10] Vasileios Lampos, Maimuna S. Majumder, Elad Yom-Tov, Michael Edelstein, Simon Moura, Yohhei Hamada, Molebogeng X. Rangaka, Rachel A. McKendry, and Ingemar J. Cox. 2021. Tracking COVID-19 using online search. *npj Digital Medicine* 4, 11 (Feb 2021), 1–11. https://doi.org/10.1038/s41746-021-00384-w

[11] Google LLC. 2020. *Google COVID-19 Search Trends symptoms dataset*.

[12] mvarch. 2022. *Multivariate Volatility Models (MVARCH) for stock prices and other time series)*.

[13] New York State Department of Health. 2022. *New York State Statewide COVID-19 Testing*.

[14] A. Rovetta. 2021. Reliability of Google Trends: Analysis of the Limits and Potential of Web Infoveillance During COVID-19 Pandemic and for Future Research. *Front Res Metr Anal* 6 (2021), 670226.

[15] K. Sato, T. Mano, A. Iwata, and T. Toda. 2021. Need of care in interpreting Google Trends-based COVID-19 infodemiological study results: potential risk of false-positivity. *BMC Med Res Methodol* 21, 1 (07 2021), 147.

[16] B. Sousa-Pinto, A. Anto, W. Czarlewski, J. M. Anto, J. A. Fonseca, and J. Bousquet. 2020. Assessment of the Impact of Media Coverage on COVID-19-Related Google Trends Data: Infodemiology Study. *J Med Internet Res* 22, 8 (08 2020), e19611.

[17] Kyunghyun Cho David Sontag Yan Liu Zhengping Che, Sanjay Purushotham. 2018. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports* (2018).

[18] Bin Zou, Vasileios Lampos, and Ingemar Cox. 2019. Transfer Learning for Unsupervised Influenza-like Illness Models from Online Search Data. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2505–2516. https://doi.org/10.1145/3308558.3313477