This is a preprint version of the article accepted by *Disaster Medicine and Public Health Preparedness*. DOI: https://doi.org/10.1017/dmp.2023.101

What Remains Now That The Fear Has Passed? Developmental Trajectory Analysis of COVID-19 Pandemic for Co-occurrences of Twitter, Google Trends, and Public Health Data

Benjamin Rathke[1]*, Han Yu[2], Hong Huang[3]

[1]Department of Applied Statistics and Research Methods, University of Northern Colorado, Greeley, CO 80639, USA, rath1045@bears.unco.edu

[2]Department of Applied Statistics and Research Methods, University of Northern Colorado, Greeley, CO 80639, USA, han.yu@unco.edu

[3]School of Information, University of South Florida, Tampa, FL, 33620, USA, honghuang@usf.edu

***Correspondent author:** han.yu@unco.edu

Abstract

The rapid onset of COVID-19 created a complex virtual collective consciousness. Misinformation and polarization were hallmarks of the pandemic in the United States, highlighting the importance of studying public opinion online. Humans express their thoughts and feelings more openly than ever before on social media, co-occurrence of multiple data sources becomes valuable for monitoring and understanding public sentimental preparedness and response to an event within our society. In this study, Twitter and Google Trends data were used as the co-occurrence data for the understanding of the dynamics of sentiment and interest during the COVID-19 pandemic in the United States from January 2020 to September 2021. Developmental trajectory analysis of Twitter sentiment was conducted using corpus linguistic techniques and word cloud mapping to reveal eight positive and negative sentiments and emotions. Machine learning algorithms were employed to implement the opinion mining how Twitter sentiment was related to Google Trends interest with historical COVID-19 public health data. The sentiment analysis went beyond polarity to detect specific feelings and emotions during the pandemic. The discoveries on the behaviors of emotions at each stage of the pandemic were presented from the emotion detection when associated with the historical COVID-19 data and Google Trends data.

Key words: social media, text mining, sentiment analysis, co-occurrence, COVID-19

Introduction

The novel coronavirus, COVID-19, impacted the daily lives and careers of millions, resulting in a flood of information and intense dialogue. Along with the public health crisis, the pandemic triggered economic and social disruption. In the United States, conversation surrounding the virus was also marred by political polarization. It is vital for governments and public health agencies to understand the nature of the public discourse surrounding COVID-19 to guide educational campaigns and inform public policy research.

Traditionally, stance has been evaluated with surveys, but there are several shortcomings (i.e., high costs, poor response rate, limited sample size, dishonest answers, and closed questions). The growing flow of information on the Internet, commonly known as Big Data, provides a new resource for meaningful insights in the digital age. Athique (2020) notes "[t]here has never been a time in which media systems have been able to convey such detailed and universal coverage of a historical event in real time, with the added capacity to keep us all in touch and to give us a voice too." Big Data, unlike survey research, relies on structuring large volumes of user-generated data. "Big Data allows us to finally see what people really want and really do, not what they say they want and say they do" (Stephens-Davidowitz, 2017, p. 54). Sources like social media and search engines have become powerful tools for analyzing real-time changes in public attitude.

Social media houses much of the sharing and consumption of news and information in the modern media environment. The demographics of users on apps like Facebook, Instagram, Twitter, and WhatsApp have historically been characterized by a younger audience, but social media platforms have lately become more representative of the general population. The past decade has seen a two-fold increase in ages 50 and older who report using at least one app ("Demographics," 2021). The growth of social media has also seen a decrease in the number of people who look to traditional media outlets for news. Two-thirds of American adults say that they "often" or "sometimes" use social media for news and about one-in-five say that it is their primary source of news (Infield, 2020; Shearer & Matsa, 2018). Twitter was a significant platform for sharing and responding to public health information and misinformation during the COVID-19 pandemic.

People on Twitter tend to be more news-focused than those on other platforms. Roughly three-quarters of Twitter users find their news on the site and two-thirds of users describe Twitter as “good” or “extremely good” for sharing health news (Shearer & Matsa, 2020). Rufai and Bunce (2020) remark that Twitter is a “powerful public health tool for world leaders to rapidly and directly communicate information on COVID-19 to citizens”. On the other hand, Shahi, et al. (2021) assert that more than four-in-five tweets may contain false claims. Due to the high volume and velocity of data production on social media, there is a reduced ability to distinguish facts from noise. Roozenbeek, et al. (2020) state that “increased susceptibility to misinformation negatively affects people's self-reported compliance with public health guidance about COVID-19, as well as people's willingness to get vaccinated against the virus and to recommend the vaccine to vulnerable friends and family.” Infield (2020) also maintains that American adults who rely on social media as their primary source of information were the most likely to believe misinformation, and the least engaged and least knowledgeable of current events. The confusing nature of information-sharing on social media may have resulted in individuals misinterpreting or disregarding public health data.

Google search data provides useful insights into understanding the discourse around COVID-19. “People’s search for information is, in itself, information” (Stephens-Davidowitz, 2017, p. 4). Google Trends measures web-based interest in topics by collating search data. “Google Trends has served and still serves as an excellent tool for infoveillance and infodemiology… newspapers and newscasts can influence web queries, it provides a way to quantify the web interest in a specific topic more efficiently than any other methods historically used (e.g., population surveys)” (Rovetta, 2021). 83% of Americans use Google as their main search engine, making Google the most popular search engine in the United States (Purcell, et al., 2012). Due to its widespread usage in the United States, web-based interest is an important factor in studying COVID-19 discourse—providing an insight into the size of the conversation about the pandemic.

With Twitter and Google Trends, a predictive model was developed for sentiment analysis with historical COVID-19 data, such as cases and deaths, through machine learning approach. With the rapid spread of misinformation during the pandemic, it remains to be unknown how COVID-19 health and policy information impacted changes in public opinion.

Literature Review

Twitter is a valuable source of big data due to its accessibility, widespread usage, availability of open-source code, and unidirectional structure (Bossetta, 2018). COVID-19 discourse has recently been examined on Twitter via frequency analysis of likes, comments and retweets, word-cloud mapping, stance detection, sentiment analysis, and network modeling (Rufai & Bunce, 2020; Tsai & Wang, 2021; Hu, et al., 2021; Fuentes & Peterson, 2021). A growing body of researchers have shown that sentiment analysis and topic modeling can be used to successfully investigate emotions and sentiment using natural language processing (Hu, et al., 2021; Schweinberger, et al., 2021; Hussain, et al, 2021; Lyu, et al., 2021). Schweinberger, et al. (2021) chose to model topics and sub-topics across different phases of the pandemic. Singh, et al. (2020) demonstrated that Twitter conversations may be used to predict the spread and outbreak of COVID-19. Hu, et al. and Hussain, et al. (2021) generated word clouds, analyzed the geo-temporal patterns of Twitter sentiment related to COVID-19, and linked changes in sentiment to key events and topics. Ahmed, et al. (2020) also generated word clouds and conducted a sentiment analysis to study the effects of lockdown and reopening procedures.

Google Trends is commonly used in conjunction with Twitter and/or health data for health research. For the MERS outbreak in 2015, Shin et al. (2016) found high correlations between the number of confirmed MERS cases and Twitter sentiment and Google interest. For the COVID-19 pandemic, Diaz and Henriquez (2021) compared Twitter sentiment and Google interest to fluctuations in the stock market and number of people under lockdown. Mavragani and Gkillas (2020) investigated the relationship between Google Trends data and COVID-19 cases and deaths. Turk et al. created a predictive model for COVID-19 cases using Google Trends and virtual consultation data. Alshahrani and Babour (2021) used Twitter and Google Trends to analyze search behaviors and predict new COVID-19 cases.

Zhang, et al., (2020) furthermore demonstrated that machine learning, specifically a unigram random forest (RF) model, is a powerful tool to predict coronavirus sentiment. Random forest regression models tend to outperform classical approaches in analyzing highly non-linear and complex relationships (James, et al., 2021). Cornelius, et al. (2021) used random forests to predict COVID-19 patient mortality. Iwendi, et al. (2020) used random forest models to predict severity of COVID-19 cases using patient geographical, travel, health, and demographic data. Random Forests are also able to produce a summary of the importance of predictors. A thorough

search of relevant literature did not yield any studies that have directly examined the effect of historical COVID-19 records (i.e., cases, deaths, vaccinations, positive tests, hospitalizations, school closures, travel bans, etc.) and Google Trends data in determining social media sentiment. Random forests are a useful tool to develop a model of using COVID-19 public health data and Google interest to predict Twitter sentiment over the course of the pandemic.

It is important to note that negative and positive events are not treated equally in public discourse. Individuals have been known to perceive negative experiences more intensely than positive ones (Rozin & Royzman, 2001; Baumeister, et al., 2001). There may be evidence that negative events are more contagious than positive events (Rozin & Royzman, 2001). On the other hand, certain key topics relating to the pandemic may be perceived more positively than expected. Yousefinaghani, et al., (2021) show that vaccine-related tweets tend to be more positive than negative. Stay-at-home tweets are also shown to be more positive than negative (Ridhwan & Hargreaves, 2021). In the context of the prolonged stress experienced by many during the pandemic, higher levels of resilience may be associated with an increase in positive emotions (Israelashvili, 2021). The complex nature of COVID-19 discourse suggests that negative sentiment may not have been the dominant emotion expressed on Twitter.

Research Questions

Q1: What were the public positive and negative sentiments on Twitter in the United States during COVID-19 pandemic?

This question is investigated by comparing the eight twitter emotion types and their dynamics over time using the data from January 1, 2020 to September 1, 2021 in the US. The exploratory study determines whether the public sentiment was evenly split between positive and negative sentiment, and that all emotions were equal, or some emotions were more common than other. For example, fear likely dominated the conversation because of the various economic, social, and health challenges experienced due to COVID-19 in the United States.

Q2: How did Google Trends and real-time historical COVID-19 data relate to sentiment on Twitter in the United States during COVID-19 pandemic?

This question is investigated by comparing twitter emotion data and google trend emotion data and their dynamics over time using data from January 1, 2020 to September 1, 2021. The

analysis examines the relationship of Google Trends and historical COVID-19 data to sentiment and emotion on Twitter over the period studied in the United States. For example, rapid increases in cases and deaths were likely significantly related to changes in sentiment and emotions on Twitter.

Data collection

Twitter data was sampled on a daily basis from January 1, 2020 to September 1, 2021 for tweets residing in the United States using the full archive search Twitter API. Zepecki, et al., (2020) outlined a methodological framework to retrieve internet data for health research, suggesting that interest be measured in respect to a list of top queries. After an exploratory analysis, Twitter and Google APIs were queried using the list of keywords “covid”, “coronavirus”, “covid19”, “corona”, “pandemic”, “quarantine”, “lockdown”, and “outbreak”. These terms were the most ~~popular~~ frequently used in discussions of COVID-19 on social media platforms. They were determined through topic analysis of all tweets over a period, as demonstrated in the studies by Schweinberger, et al. (2020) and Hu, et al. (2021). Future studies may first do a relevant topic analysis, then pull relevant tweets for a more representative sample. A unigram (one-word) method was chosen because of its optimal use in random forest models (Zhang, et al., 2020). 2,500,000 tweets were pulled, and just under 900,000 unique tweets were identified for this study.

Shortly after COVID-19 was discovered, there was little discussion about the virus. Some days therefore have a small number of tweets which leaves the subsequent analysis vulnerable to sampling error. To avoid this, sampling was constructed at three locations throughout each day as outlined by Kim et al. (2018). Geo-tweet information is provided when users activate location access and provide a finer geographical scale, however not all users activate this function. According to Twitter, only 30-40% of tweets contain information about profile location (“Advanced Filtering”). It was deemed that geographical analysis was not generalizable enough, so state-level and city-level granularity was not included in this study. Tweets were preprocessed to remove retweets, references to screen names, hashtags, spaces, numbers, punctuations, urls, retweet headers, time codes, stop-words, and duplicate tweets.

Google Trends data was obtained using the Trends API and gtrendsR endpoint in R. Google Trends returns data in daily granularity only if the timeframe is shorter than 9 months, so daily estimates for each month and monthly data for the entire time frame were retrieved, and daily estimates for each month were multiplied by the weight calculated from monthly data to calculate daily estimates from January 1, 2020 to September 1, 2021. Google Trends estimated interest is shown in Figure 1.

Historical data about the virus was supplied by Our World in Data from the COVID-19 Data Repository by the Center for Systems Science and Engineering at John Hopkins University, government sources, and peer reviewed research. This dataset includes confirmed cases, confirmed deaths, vaccinations, hospital and ICU, tests and positivity, the reproduction rate of the virus, policy responses, and other variables of interest. Missing data was substituted with estimated values from near neighbors as outlined by Kang (2013). New cases and new deaths over time are visualized in Figures 2 and 3 respectively.

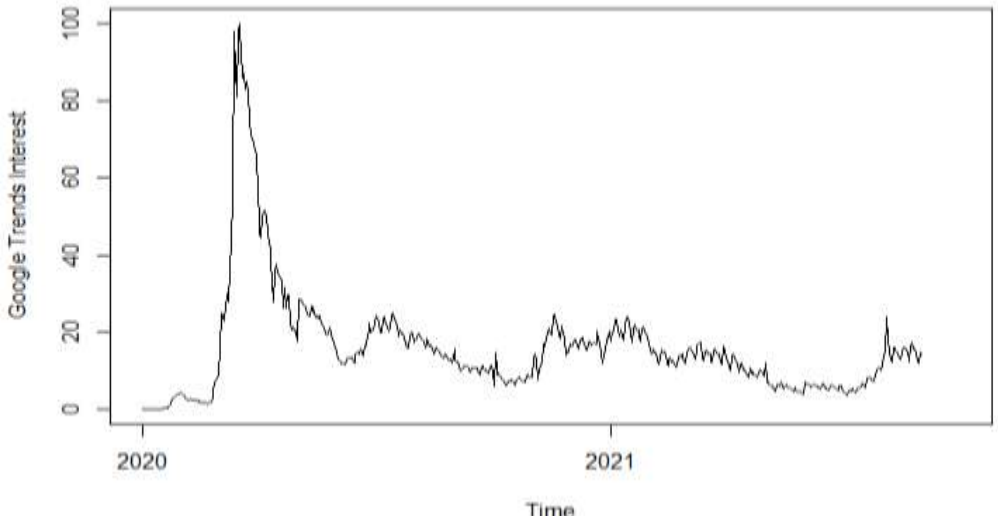


Figure 1 Google Trends interest over time in USA

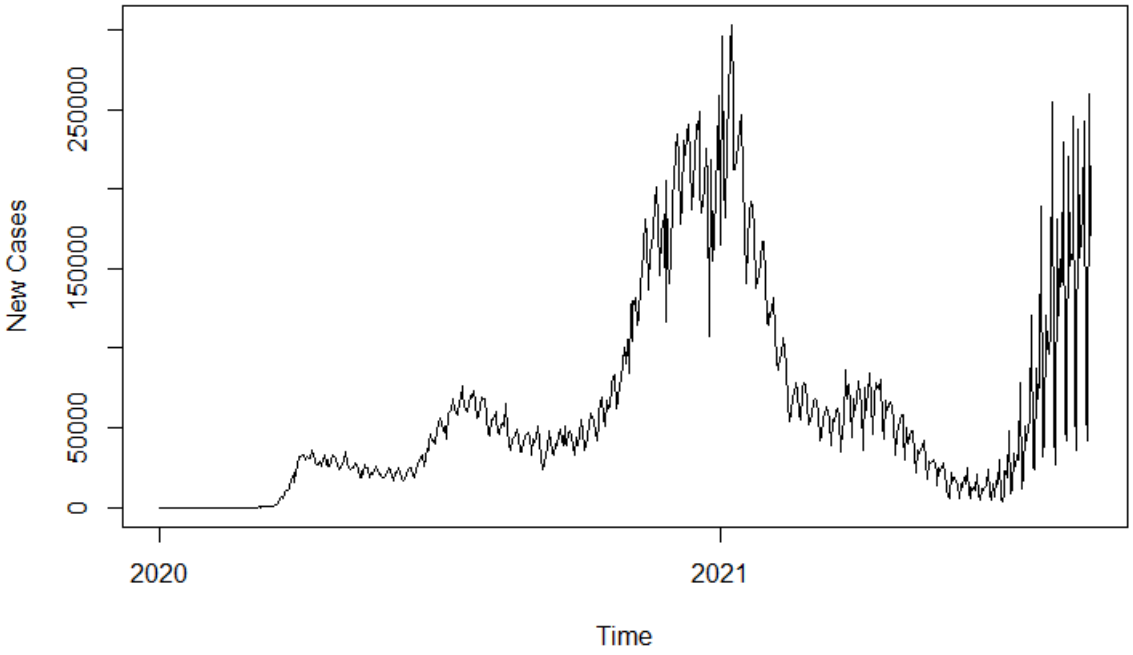


Figure 2: New COVID-19 cases over time in USA

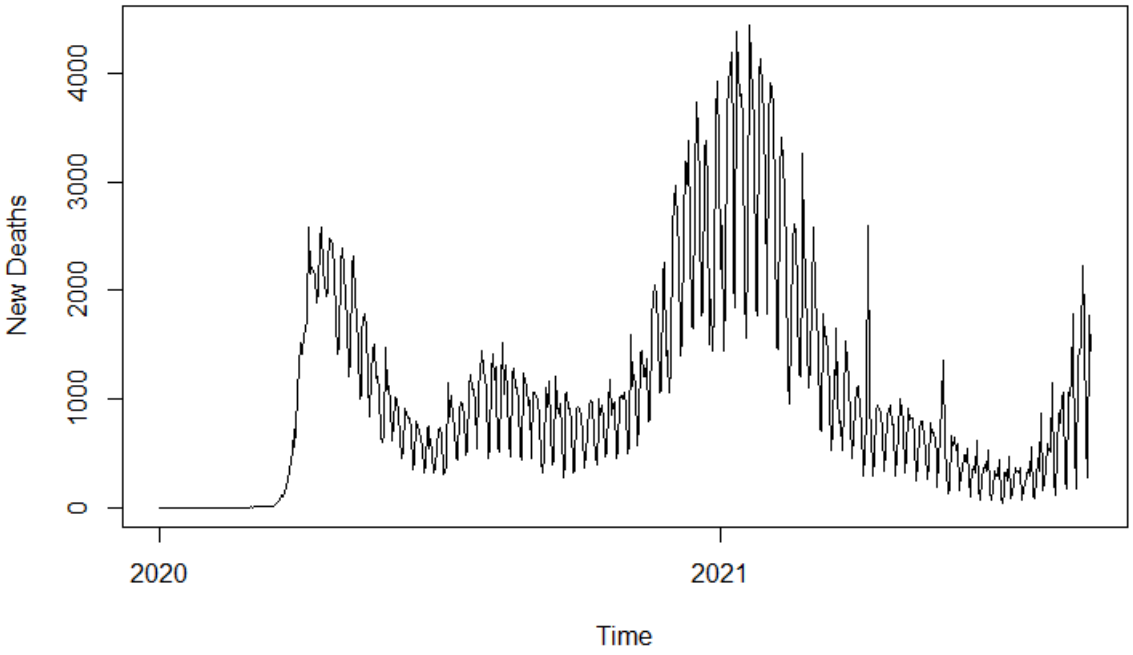


Figure 3: New COVID-19 deaths over time in USA

Data Summary

The summary statistics of each variable included in this study, including historical COVID-19 health and policy data, Twitter sentiment (positive, negative, trust, surprise, sadness, joy, fear, disgust, anticipation, anger), and Google Trends interest are given in Table 1. Note that vaccinations and boosters contained many null values because vaccines were only available later in the pandemic.

Table 1: Summary Statistics

	Min	1st Qu	Median	Mean	3rd Qu	Max
Date	2020-1-1	2020-6-1	2020-10-31	2020-10-31	2020-4-1	2021-8-31
total cases	0	1816679	9167578	15334936	30593758	39321999
new cases	0	21735	44345	64568	75008	303008
total deaths	0	108445	231515	296881	552577	640859
new deaths	0	352	845	1052	1440	4441
total cases per million	0	5457	27537	46063	91897	118114
new cases per million	0.00	65.3	133.2	194.0	225.3	910.2
total deaths per million	0.0	325.7	695.4	891.8	1659.8	1925.0
new deaths per million	0.0	1.1	2.5	3.2	4.3	13.3
reproduction rate of covid	0.00	0.87	1.00	1.02	1.14	3.65
icu patients	0	0	8419	9034	13479	28891

icu patients per million	0.0	0.0	25.3	27.1	40.5	86.8
hospitalized patients	0	0	30283	35452	47215	133253
hospitalized per million	0.0	0.0	91.0	106.5	141.8	400.3
new tests	0	414634	868519	877732	1326296	2323355
total tests	0	19696742	155419699	205510660	386391223	534538919
new tests per thousand	0.00	1.25	2.61	2.64	3.98	6.98
total tests per thousand	0.00	59.16	466.85	617.31	1160.63	1605.63
positive rate	0.000	0.042	0.054	0.063	0.092	0.206
tests per case	0.0	9.3	15.6	16.5	21.9	56.0
total vaccinations	0	0	0	83213272	153631404	370212027
people vaccinated	0	0	0	48447862	99565311	205026070
people fully vaccinated	0	0	0	36759492	56089614	174121529
total boosters	0	0	0	8287	0	995715
new vaccinations	0	0	0	607443	990875	4629928
total vaccinations per hundred	0.0	0.0	0.0	24.7	45.7	110.1
people vaccinated per hundred	0.0	0.0	0.0	14.4	29.6	61.0

people fully vaccinated per hundred	0.0	0.0	0.0	10.9	16.7	51.8
government stringency index (policy response)	0.00	56.02	65.28	57.34	71.76	75.46
absolute cumulative excess mortality	-13814	121838	288041	339819	614172	720824
anger (Twitter)	8	492	570	555	642	942
anticipation (Twitter)	34	697	794	786	944	1232
disgust (Twitter)	8	373	428	424	492	767
fear (Twitter)	21	773	896	919	1071	1597
joy (Twitter)	19	490	562	548	671	954
sadness (Twitter)	15	619	705	695	798	1227
surprise (Twitter)	11	377	437	426	498	1355
trust (Twitter)	27	797	924	912	1079	1568
negative (Twitter)	33	1196	1353	1344	1535	2253
positive (Twitter)	50	1261	1468	1460	1758	2502
est. interest (Google Trends)	0.087	7.617	13.797	16.060	19.194	100

Methodology

Corpus-linguistic techniques were used to create a word cloud of most used words in sampled tweets. The National Research Council Lexicon dictionary (NRC-Lex) was used to conduct sentiment analysis. The NRC-Lex dictionary is based on the eight emotion classifications (joy, sad, anger, fear, trust, disgust, surprise, anticipation) and sentiment (positive or negative). Frequencies of each emotion and sentiment were obtained in time series.

Sentiment prediction was achieved using random forest (RF) models. Twitter sentiment counts, Google Trends estimated interest, and historical COVID-19 data was aggregated by day, and ten Random Forest models were developed for each sentiment type. A training dataset was formed with two-thirds of the data, and a test set was formed with the remaining rows. Mean absolute percentage error (MAPE) was calculated for training and test sets. Important parameters can be calculated for random forest models based on node purity and minimal depth. Both indexes are effective, but node purity was chosen as the primary method for this study. Unimportant variables were discarded to prevent overfitting, and a new model was appropriately refitted for each sentiment type using the most important variables. Including relevant variables improves the performance of random forests.

Random Forests

Random forests (Breiman, 2001) are a substantial modification of bootstrap aggregation (bagging), a variance-reduction technique for an estimated predictive function through building a large collection of de-correlated trees with each generated tree being identically distributed and averages the resulting trees. Trees are ideal candidates for sentiment analysis since they can capture complex interaction structures inherent in the highly correlated text data. Trees have relatively low bias if grown sufficiently deep. However, trees are notoriously noisy and thus need averaging. Using stochastic perturbation and growing and averaging trees on samples avoid overfitting. The algorithm is as follows in Table 2,

Table 2. Algorithm of Random Forest

1. For b = 1 to B:
(I) Draw a bootstrap sample Z^* of size N from the training data. (II) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached. i. Select m variables at random from the p variables. ii. Pick the best variable as split-point among the m. iii. Split the node into two daughter nodes.
2. Output the ensemble of trees {T .
3. (x)=– (x).

After Bth recursion, tree sequences Θ are grown, the random forest predictor at a single target point x is

(x)=– Θ),

where Θ parameterizes the bth random forest tree in the sequence in terms of split variables, cutpoints at each node, and terminal-node values.

Random forests cannot overfit the data. Increasing B does not cause the random forest to overfit as

(x)=$E_{\Theta|}$ T(x; Θ(Z))=lim $_{\infty}$

with an average over B realizations of Θ(Z) and the distribution of Θ(Z) is conditional on the training data Z. Using full-grow trees result in one less tuning parameter and seldom costs much. The robustness is largely due to the relative insensitivity of misclassification cost to the bias and

variance of the probability estimates in each tree. Let ρ(x) is the conditional sampling correlation between any pair of trees used in the averaging,

ρ(x)=cor(T(x; Θ (Z), T(x; Θ (Z)),

where Θ (Z) and Θ (Z) are a randomly drawn pair of random forest tree grown to the randomly sampled Z. (x) is the sampling variance of any single randomly drawn tree, (x)=Var(T(x; Θ(Z)).

Then the

Var(f (x))= ρ(x) (x).

The conditional covariance of a pair of tree fits at x is zero due to the fact that the bootstrap and feature sampling is independent and identically distributed (i.i.d). On many problems the performance of random forests is very similar to boosting, and they are simpler to train and tune. Hastie et al. (2016) made grand claims that random forests are "most accurate", "most interpretable", and the like with very little tuning required.

Sentiment Analysis

To address the first research question, frequency counts from the sentiment analysis of sampled tweets using the terms "covid", "coronavirus", "covid19", "corona", "pandemic", "quarantine", "lockdown", and "outbreak" were totaled independent of time to produce the findings in Figure 4. Figure 4 shows that, over the course of the period studied, sentiment tended to be more positive than negative. Fear was the most popular emotion, followed closely by trust. Other emotions were less common, including anticipation, sadness, anger, joy, surprise, and disgust.

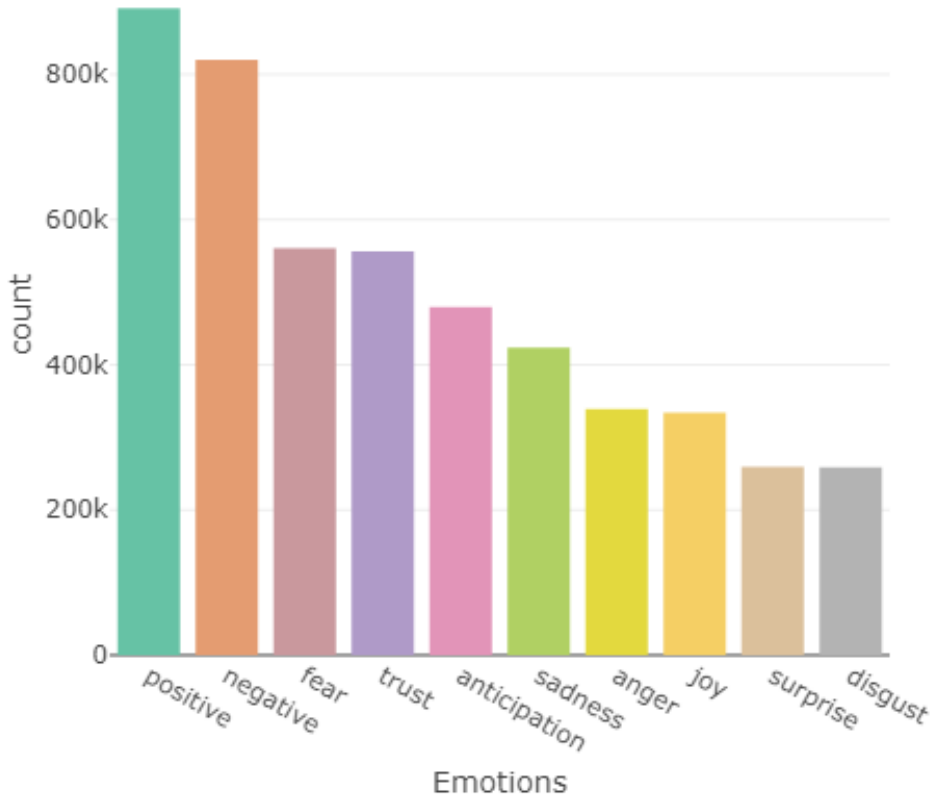


Figure 4: Emotion type for COVID-19 tweets

Another perspective on sentiment is given with the word cloud in Figure 5, which shows the most popular words in the Twitter sample. The most popular words were “quarantine” and “trump”. Figure 5 also portrays how words were associated with emotions from the sentiment analysis. Note, this word cloud was weighted toward the keywords that were used, and did not include all popular words due to spacing constraints.

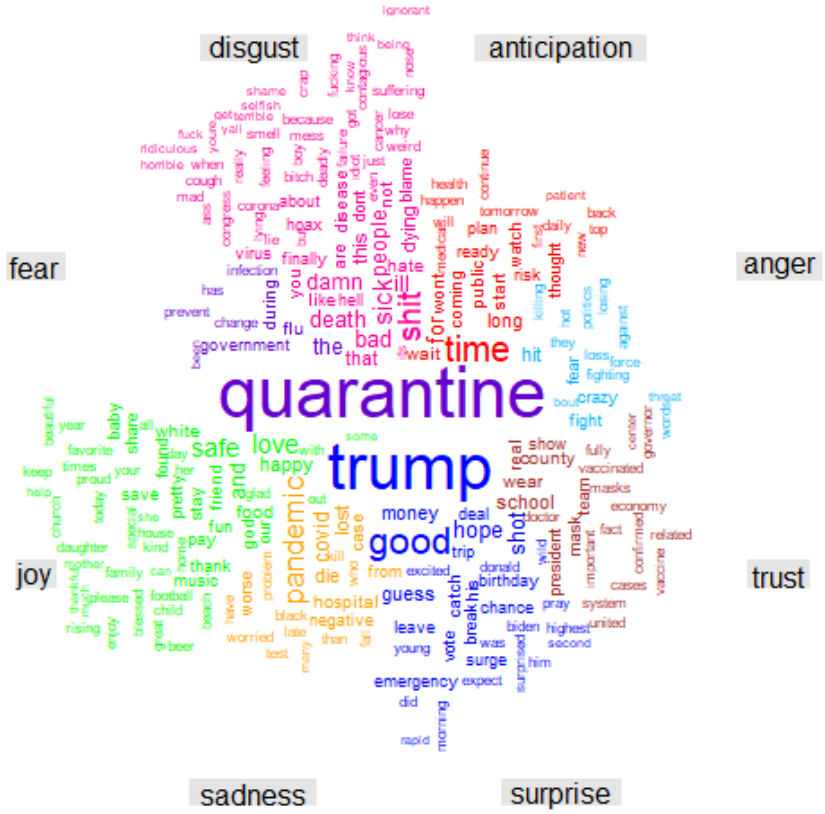


Figure 5: Word Cloud for emotional terms in

The temporal trajectories of observed and predicted sentiment are also plotted over the time. Figure 6 used the complete dataset, Figure 7 used the training dataset for the random forest (420 observations), and Figure 8 used the predicted values of each random forest model. Green signifies positive sentiment, while red is negative sentiment. The other colors – purple, orange, glue, aquamarine, chartreuse, black, yellow, and pink – correspond to trust, surprise, sadness, joy, fear, disgust, anticipation, and anger respectively. Notably, all sentiment types tended to follow similar trends.

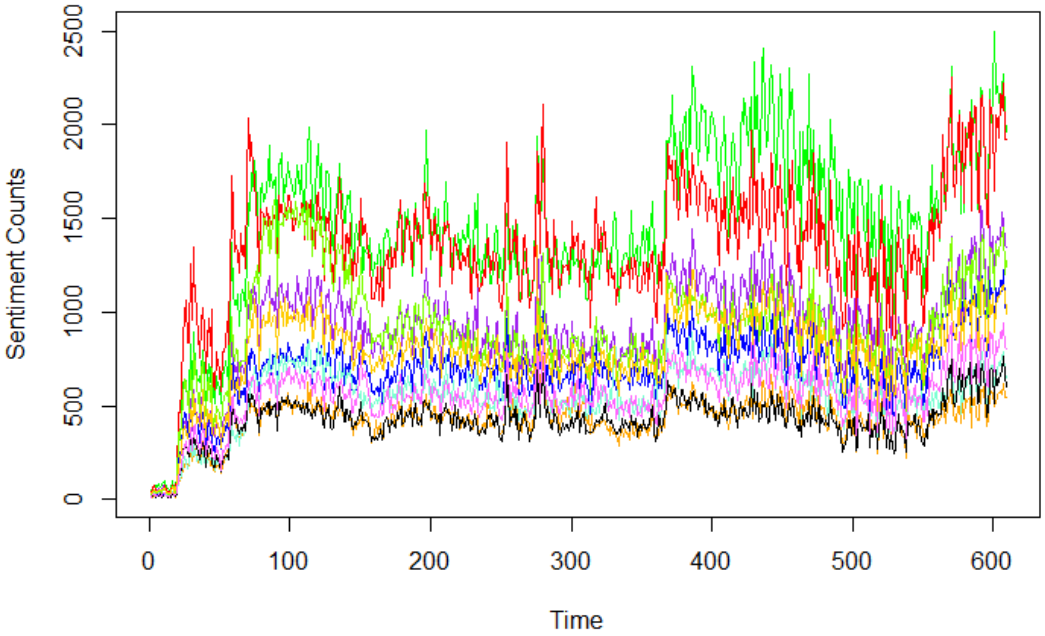


Figure 6: Trajectories of observed sentiment counts over time in USA

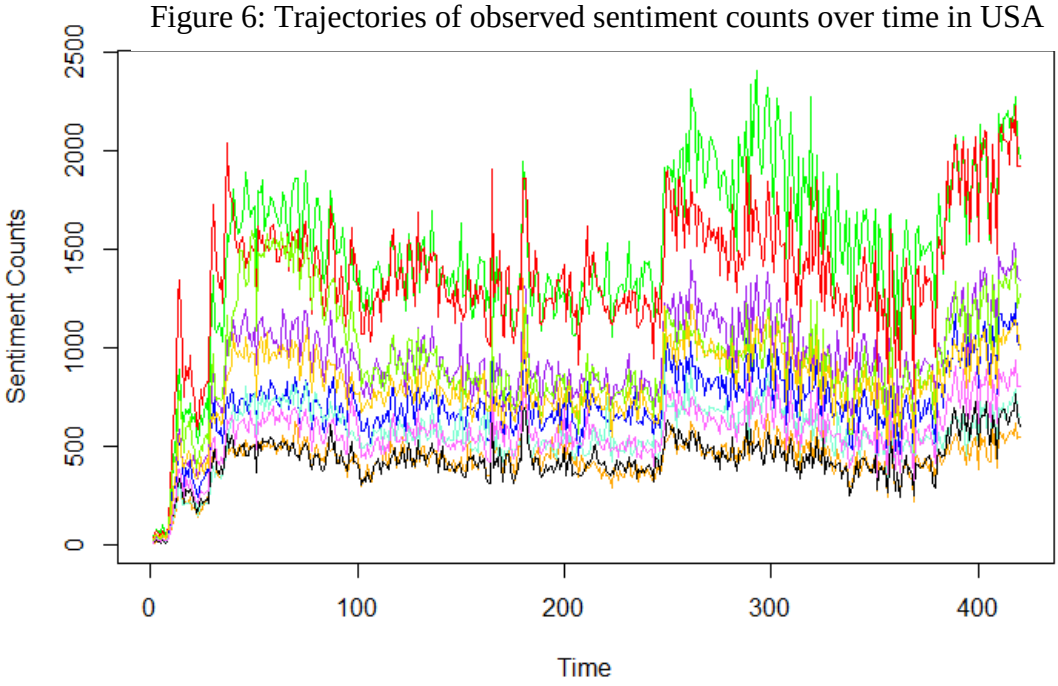


Figure 7: Trajectories of training data from observed sentiment counts over time in USA

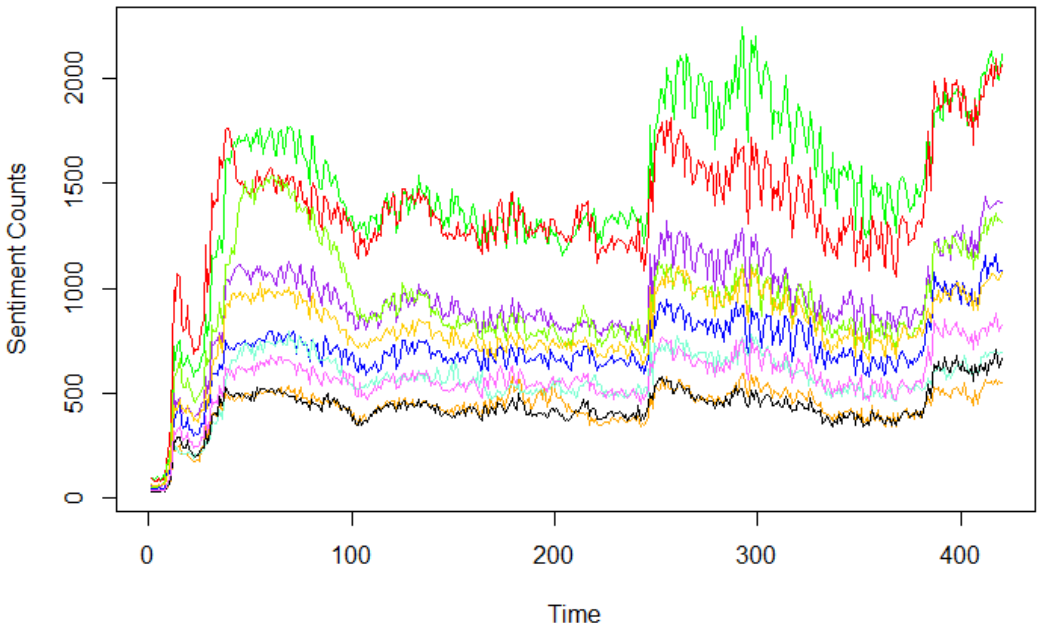


Figure 8: Trajectories of predicted sentiment counts over time in USA

Visually, it appears that the predictive models performed quite match with the actual data. The Mean Absolute Percentage Error (MAPE) and the reported percentage of variation explained quantified how well the random forest models fit the data. MAPE was produced for both the training and the test sets to investigate overfitting and generalizability in Table 2. A MAPE score of less than 20% was considered excellent, while scores from 20% to 30% were considered good. The MAPE for the test set was consistently two to three times higher than the training set indicating overfitting, however the MAPEs for all training and test sets had relatively low values. Additionally, the percentage of variation explained was adequate for all models. The surprise sentiment model performed the worst.

Table 2: Random Forest model performance for sentiment types

Sentiment Type	MAPE Training Set	MAPE Test Set	%Var explained
Positive	4.60%	10.41 %	84.98%
Negative	5.16%	12.61%	79.13%
Trust	4.84%	11.53%	83.04%
Surprise	5.97%	14.73%	65.07%
Sadness	5.59%	15.16 %	81.26%
Joy	4.63%	10.82%	80.51%
Fear	4.74%	12.17%	86.77%
Disgust	6.47%	15.61%	76.8%
Anticipation	4.43%	10.65%	81.44%
Anger	5.67%	14.56%	79.21%

The important parameters for each random forest model are now detailed for each sentiment type with plots of observed and predicted sentiment provided for reference.

Results

Positive Sentiment Random Forest

The significant parameters for positive sentiment random forest are shown in Figure 9. As a proof of concept, minimum depth important parameters and interaction important are also plotted to compare important parameters decided by node purity and are shown in Figure 10 and 11. With cross validation of the mean of minimal depth distribution and interaction (Figure 10 and 11), among all the parameters, “date”, “total_cases_per_million”, “total_cases” and “est_hits” are important ones for positive sentiment random forest.

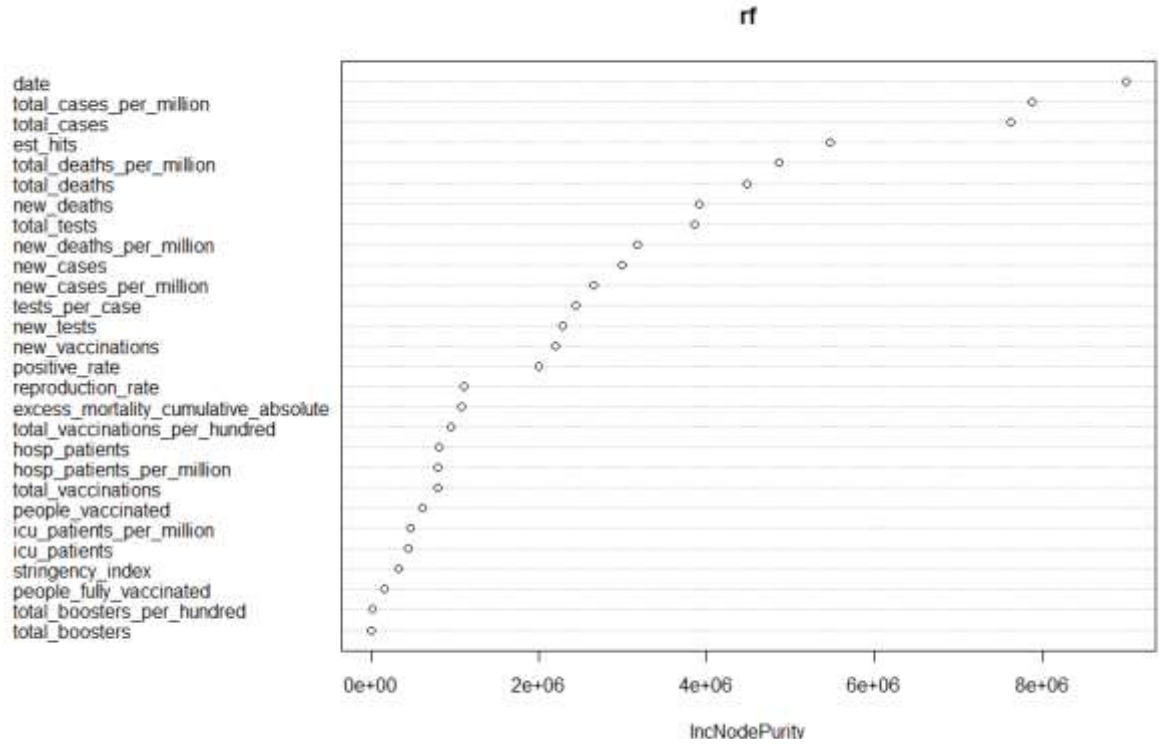


Figure 9: Important parameters for positive sentiment random forest using node purity

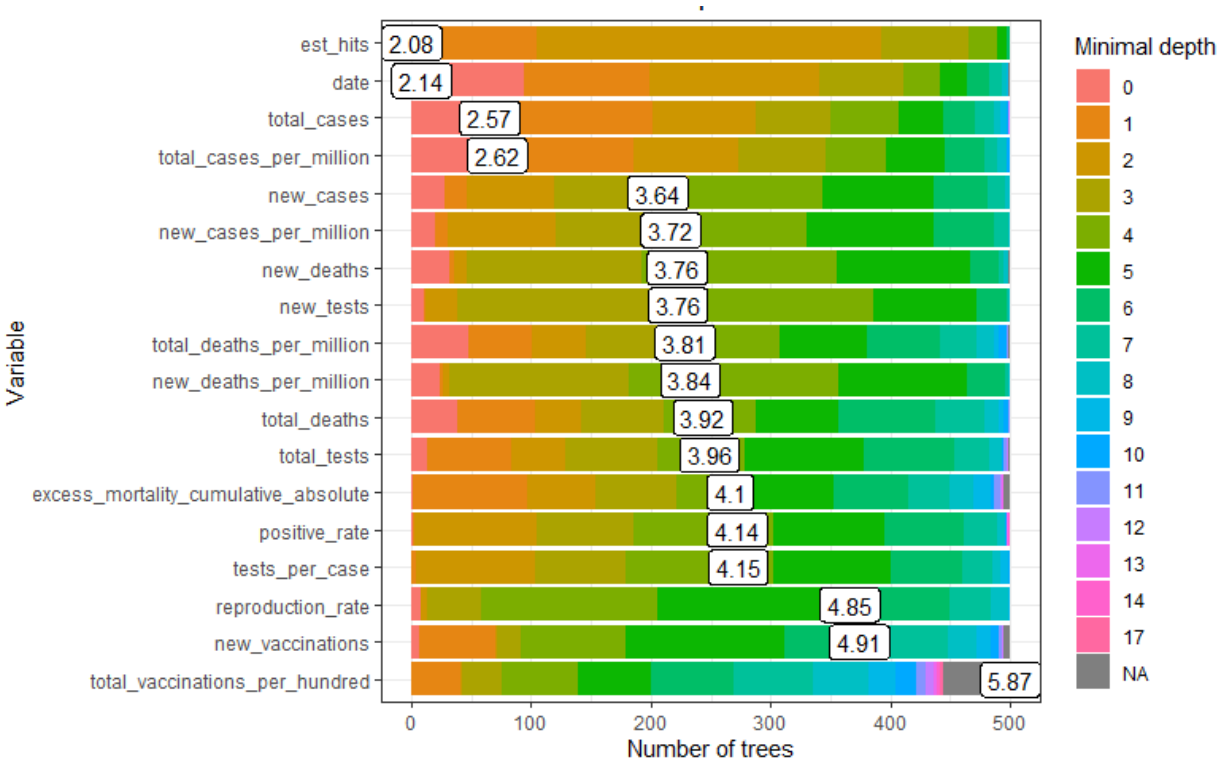


Figure 10: Important parameters for positive sentiment random forest using distribution of minimal depth and its mean

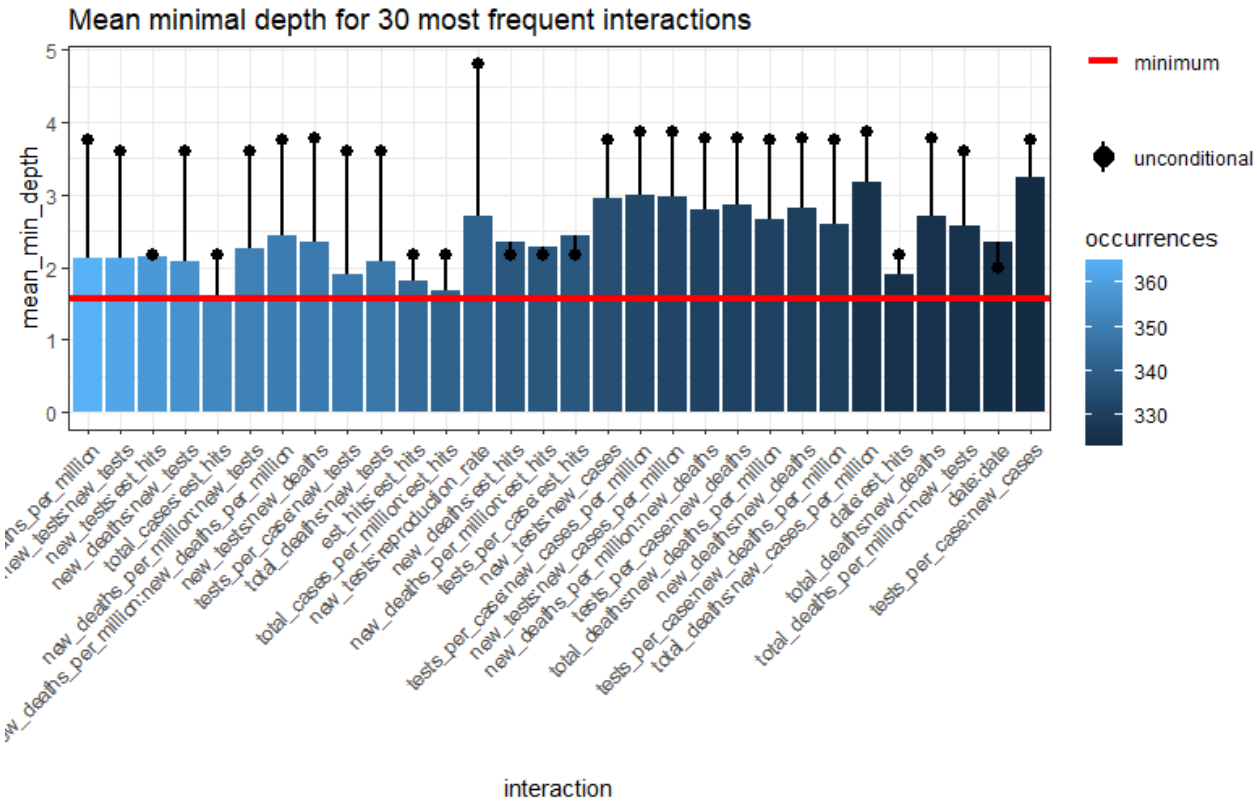


Figure 11: Important parameters for positive sentiment random forest using interaction

Node purity and minimal depth provided similar results for deciding important parameters. Interaction methods were deemed too complex to interpret and were not used for the analysis. The observed and predicted positive sentiments over time are shown in Figure 12. Positive sentiment increased during the starting of pandemic, then was stable later, another wave observed starting in 2021.

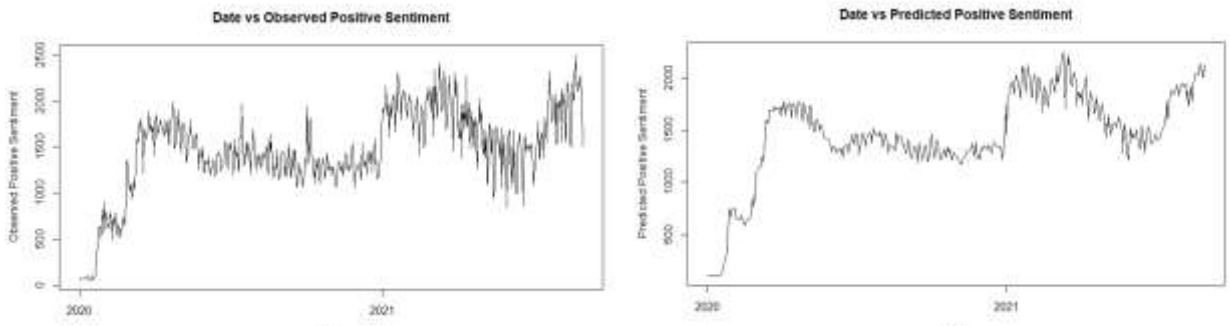


Figure 12: Observed (left) and predicted (right) positive sentiment vs time

Negative Sentiment Random Forest

The significant parameters for negative sentiment random forest are shown in Figure 13. "est_hits", "date", "total_cases", and "total_cases_per_million" are the important parameters for negative sentiment random forest. Notably, Google Trends interest appears to be the most important variable for prediction.

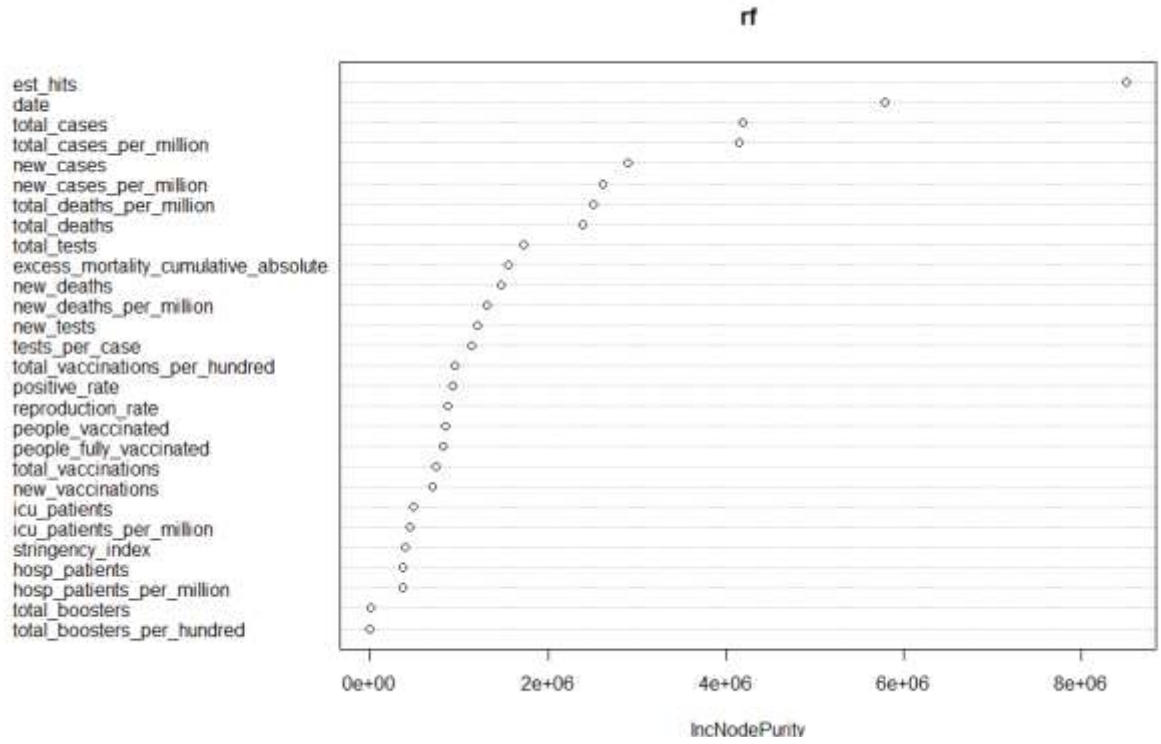


Figure 13: Variable important plot for negative sentiment random forest using node purity

The observed and predicted negative sentiment over time are shown in Figure 14. The negative sentiments increased at the beginning of the COVID19, with fluctuation over time.

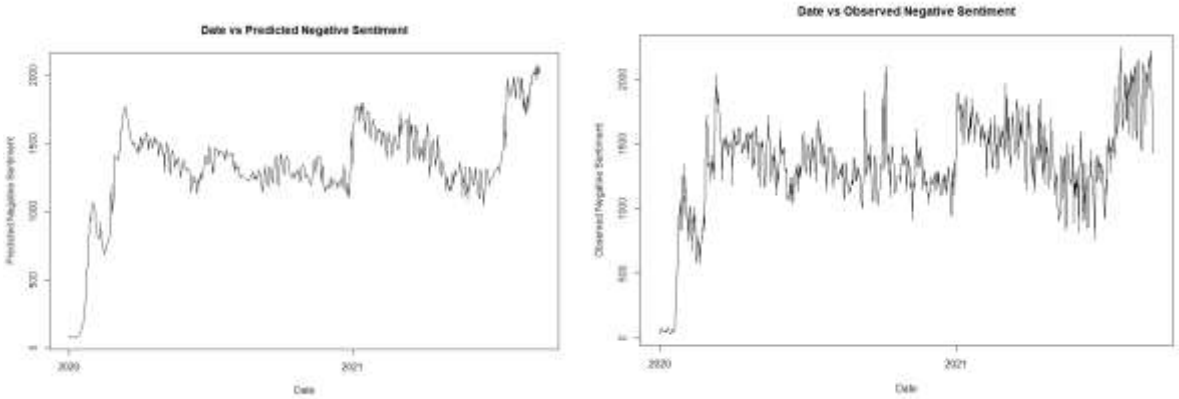


Figure 14: Observed (left) and predicted (right) negative sentiment vs time

Positive and negative emotions exhibit distinct trend patterns over time (see Appendix). Sentiment frequency over time diagrams were redrawn to better illustrate trend patterns. All positive sentiments, including Trust, Surprise, Joy, and Anticipation, dramatically increased at the start of COVID19 in 2020, and fluctuate over time, with a second peak observed at the start of 2021, but the overall shape is flat (Figure 15). Nonetheless, negative emotions such as Sadness, Anger, and Disgust increased rapidly at the start of the pandemic, with a minor drop later, and then remained stable with a degree of fluctuation, before continuing to rise and reaching a peak in late 2021 (Figure 16). Interestingly, fear sentiment appears in the first wave at the start of COVID19, then falls noticeably, and then returns with a spike at the end of 2021, but at a lower level than the initial jump (Figure17).

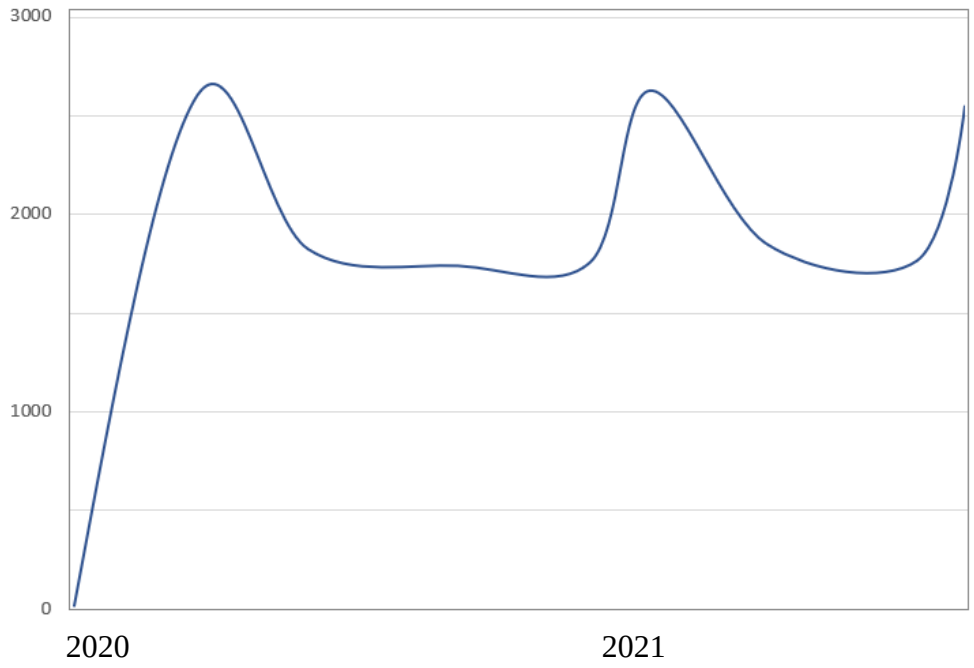


Figure 15 Sentiment trend patterns over time for positive emotion: Trust, Surprise, Joy and Anticipation

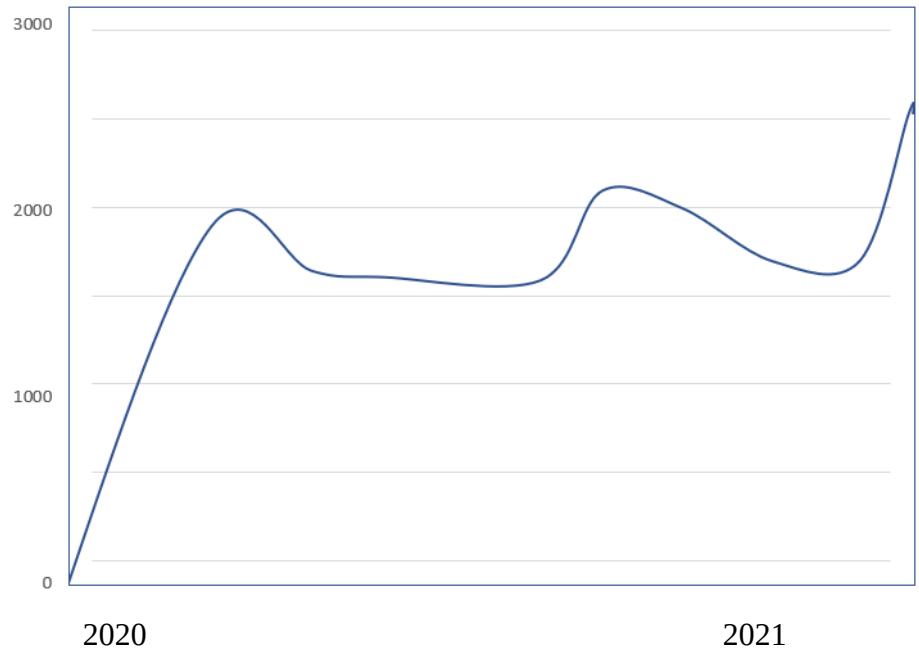


Figure 16. Sentiment trend over time patterns for negative emotion: Sadness, Disgust, and Anger

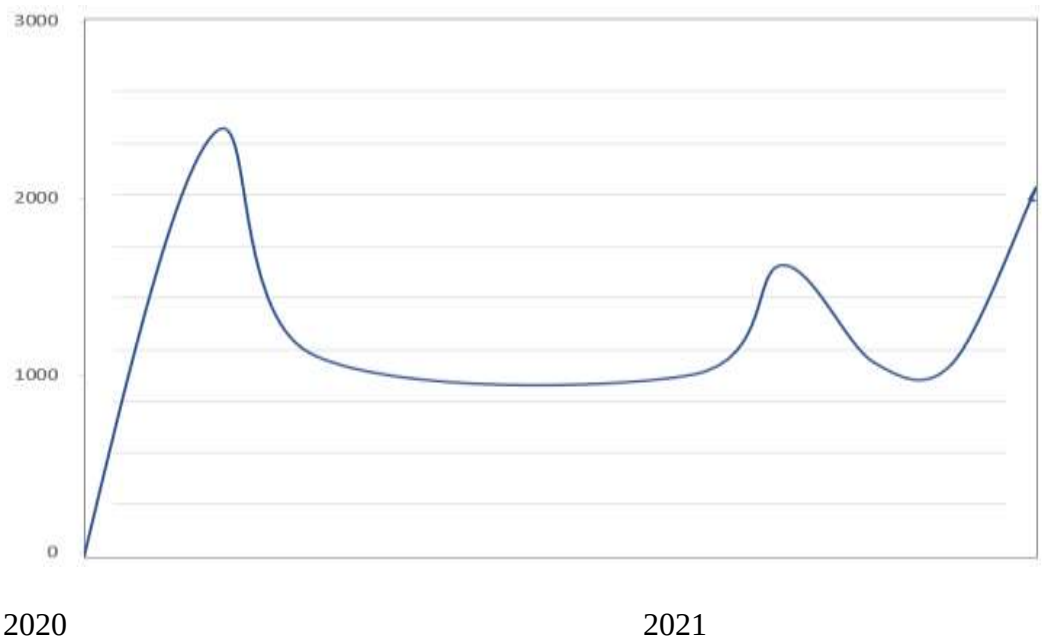


Figure 17. Sentiment trend pattern over time for negative emotion:Fear over time

Discussion and conclusions

The number of people using social media platforms and search engines has increased dramatically during the digital age. The consumption of news on social media has grown, bringing both lower engagement and a diminished understanding of current events. In the United States, the internet became a significant source of misinformation during COVID-19 amid social, economic, and public health crises. Twitter and Google Trends provide valuable insights into public discourse surrounding COVID-19. This study presented the results of a sentiment analysis of tweets, Google Trends interest, and historical COVID-19 health and policy data over the course of the pandemic and built a predictive model for sentiment.

Sentiment analysis revealed that people mentioned “quarantine” and “Trump” the most. These were some of the most important topics during the pandemic; however, they were weighted toward the keywords in the tweet sample. For example, “quarantine” may not have been as important as the word cloud represented because it was also one of the keywords used to find relevant tweets. Positive sentiments were more common than negative sentiments, while fear and trust were the most common emotions. The sentiment analysis in the present study agreed with Hu et al. (2021), Hussain et al. (2021), and Ahmed et al. (2020).

Google Trends interest showed a sharp peak at the beginning of the pandemic, which seemed to be related to the first peaks in COVID-19 cases and deaths. This indicates that people in the United States searched for COVID-19 primarily at the beginning of the pandemic as cases and deaths were first appearing. Google Trends estimated interest agreed with analyses by Mavragani and Gkillas (2020), Turk et al. (2021), and Alshahrani and Babour (2021).

Random forest models were used to predict sentiment types. The most important factors for all models were date, COVID-19 cases, COVID-19 deaths, and Google Trends estimated interest. These models showed that Google Trends and public health data were both important indicators for changes in sentiment. For positive sentiment, the most important factor was date, but for negative sentiment, the most important factor was Google Trends interest. This makes sense given the relationship of Google Trends interest to COVID-19 cases and deaths. The number of people vaccinated did not affect sentiment as much as the number of cases or deaths. Vaccinations were undervalued in the present analysis – due to the large time range there are too many zero values to notice an effect. It is worth noting that for fear and joy sentiments, COVID-19 tests were also an important variable. Positive emotions during COVID19 might be explained

to link to the recovery progress, vaccine development, new hopes of technologies development, and resilience (Israelashvili, 2021).

Anger, disgust and sadness sentiments keep up increasing during the pandemic, indicating people in the U.S. emotionally are unexpected for such a long duration of the pandemic. Fear is easy come and easy go. Fear sentiment shows a big wave at the beginning of COVID19 since 2020, later on drops gradually, at last, have a big jump at the end of 2021. Fear sentiment cannot last long, but if the event is persistent, it will come back later. Joy is a kind of positive sentiment, like positive sentiment demonstrates a flat and wavy behavior, reflecting a hope at the beginning of 2020 when COVID19 starts, and at the beginning of 2021. Anticipation, surprise and negative sentiments show a series of fluctuation waves. This indicated that people gave thoughts and analysis, and information seeking behaviors reflected by Google Trends interest.

However, there were several limitations. Twitter tends to represent a younger audience, and does not include the entire conversation surrounding COVID-19. In addition, elderly, poor, and underprivileged members are underrepresented on the internet. More work needs to be done to smooth the noise in sentiment scores. The present analysis only accounts for the keywords used to query Twitter and Google, and do not represent all possible topics. For a more representative sample, we may have sampled from all available tweets/searches and identified those that were related to COVID-19 using topic analysis. Future research may also use a different sentiment/emotion database to acquire a more diverse look than the 10 sentiment types in this study.

In this study, “vaccine(s)” was not included for key word search. Sentiment related to vaccines is an important aspect of the public's perception of the pandemic, as the widespread availability and acceptance of vaccines is seen as key to controlling the spread of the virus and eventually bringing the pandemic to an end. However, the decision was made not to include vaccines in queries to maintain a clear interpretation of the relationships between overall sentiment of COVID-19 on Twitter and the predictors. Vaccine sentiment may have introduced nuanced correlations in the presence of misinformation and politics. Future study would conduct a topic analysis in depth to identify terms relating to COVID-19 and stratify keywords into sub-topics including vaccines.

The current research focus on emotion analysis from text at this stage because text is still the primary choice for people to express their feelings toward other persons, events, or things. However, a multi-platform approach, such as using CrowdTangle, for richer sources of information can be valuable to analyze emotions. A multi-platform approach may have provided a more comprehensive view of public sentiment. For future research, we will consider incorporating data from additional platforms for the analysis in context for the data noise such as sarcasm and irony.

Extracting emotions behind text is still an immense and complicated task in current literature. The study contributes to existing literature by directly examining the effect of health data and Google Trends interest to Twitter sentiment over the duration of the pandemic. The information from this study can be used to acquire a better understanding of COVID-19's emotional impact on people and communities, as well as their fears, concerns, and coping mechanisms. Furthermore, tracking the emotional patterns of COVID-19-related tweets over time can offer a more thorough picture of how public views and perceptions of the pandemic are changing. Overall, monitoring COVID-19-related tweets for emotion change can support public health research and help inform strategies to address the impacts of the pandemic on individuals and communities.

Conflict of Interests

The authors declare that there is no conflict of interests.

References

"Advanced Filtering for Geo Data | Docs | Twitter Developer Platform." *Developer Platform*, Twitter, https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data.

"Demographics of Social Media Users and Adoption in the United States." *Pew Research Center*, Washington, D.C. 2021, https://www.pewresearch.org/internet/fact-sheet/social-media/?menuItem=b14b718d-7ab6-46f4-b447-0abd510f4180.

Ahmed, Mohammed Emtiaz, et al. "Covid-19: Social Media Sentiment Analysis on Reopening." *ArXiv.org*, 1 June 2020, https://arxiv.org/abs/2006.00804.

Alshahrani, Reem, and Amal Babour. "An Infodemiology and Infoveillance Study on COVID-19: Analysis of Twitter and Google Trends." *Sustainability*, vol. 13, no. 15, 30 July 2021, p. 8528., https://doi.org/10.3390/su13158528.

Al-Zaman, M. S. (2021). "Prevalence and Source Analysis of COVID-19 Misinformation in 138 Countries." *IFLA Journal*, 2021, https://doi.org/10.1177/03400352211041135

Athique, A. (2020). "Extraordinary Issue: Coronavirus, Crisis and Communication." *Media International Australia*, vol. 177, no. 1, 2020, pp. 3–11, https://doi.org/10.1177/1329878X20960300

Barrie, Christopher and Ho, Justin Chun-ting. (2021). "academictwitteR: an R package to access the Twitter Academic Research Product Track v2 API endpoint." *Journal of Open Source Software*, 6(62), 3272, https://doi.org/10.21105/joss.03272

Bossetta, Michael. "The Digital Architectures of Social Media: Comparing Political Campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election." *Journalism & Mass Communication Quarterly*, vol. 95, no. 2, 2018, pp. 471–496., https://doi.org/10.1177/1077699018763307.

Breiman, L. (2001). Random forests, Machine Learning 45: 5-32.

Cornelius, Erwin, et al. "COVID-19 Mortality Prediction Using Machine Learning-Integrated Random Forest Algorithm under Varying Patient Frailty." *Mathematics* vol. 9, no. 17, 2021, pp. 2043., https://doi.org/10.3390/math9172043

Díaz, Fernando, and Pablo A. Henríquez. "Social Sentiment Segregation: Evidence from Twitter and Google Trends in Chile during the COVID-19 Dynamic Quarantine Strategy." *PLOS ONE*, vol. 16, no. 7, 2021, https://doi.org/10.1371/journal.pone.0254638.

Fuentes, Agustín, and Jeffrey V. Peterson. "Social Media and Public Perception as Core Aspect of Public Health: The Cautionary Case of @realdonaldtrump and Covid-19." *PLOS ONE*, vol. 16, no. 5, 2021, https://doi.org/10.1371/journal.pone.0251179.

Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" 2nd Edition. Springer. 2016

Hu, Tao, et al. "Revealing Public Opinion towards Covid-19 Vaccines with Twitter Data in the United States: a Spatiotemporal Perspective." *Journal of Medical Internet Research*, vol. 23, no. 9, 2021, https://doi.org/10.2196/30854.

Hughes, Adam, and Stefan Wojcik. “10 Facts about Americans and Twitter.” *Pew Research Center*, Washington, D.C. 16 Sept. 2020, https://www.pewresearch.org/fact-tank/2019/08/02/10-facts-about-americans-and-twitter/.

Hussain, Amir, et al. “Artificial Intelligence–Enabled Analysis of Public Attitudes on Facebook and Twitter toward Covid-19 Vaccines in the United Kingdom and the United States: Observational Study.” *Journal of Medical Internet Research*, vol. 23, no. 4, 5 Apr. 2021, https://doi.org/10.2196/26627.

Infield, Tom. “Americans Who Get News Mainly on Social Media Are Less Knowledgeable and Less Engaged.” *The Pew Charitable Trusts*, Washington, D.C. 16 Nov. 2020, https://www.pewtrusts.org/en/trust/archive/fall-2020/americans-who-get-news-mainly-on-social-media-are-less-knowledgeable-and-less-engaged.

Israelashvili, J. (2021). More positive emotions during the COVID-19 pandemic are associated with better resilience, especially for those experiencing more negative emotions. *Frontiers in Psychology*, *12*, 1635.

Iwendi, Celestine, et al. “COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm.” *Frontiers in Public Health*, vol. 8, no. 357, 3 July 2020, https://doi.org/10.3389/fpubh.2020.00357.

James, Gareth, et al. “Tree-Based Methods.” *An Introduction to Statistical Learning: with Applications in R*, 2nd ed., Springer, New York, NY, 2021.

Kang, Hyun. “The Prevention and Handling of the Missing Data.” *Korean Journal of Anesthesiology*, vol. 64, no. 5, 24 May 2013, pp. 402–406., https://doi.org/10.4097/kjae.2013.64.5.402.

Kim, Hwalbin, et al. “Evaluating Sampling Methods for Content Analysis of Twitter Data.” *Social Media + Society*, vol. 4, no. 2, 2 May 2018, https://doi.org/10.1177/2056305118772836.

Lyu, Joanne Chen, and Garving K Luli. “Understanding the Public Discussion about the Centers for Disease Control and Prevention during the COVID-19 Pandemic Using Twitter Data: Text Mining Analysis Study.” *Journal of Medical Internet Research*, vol. 23, no. 2, 9 Sept. 2021, https://doi.org/10.2196/25108.

Mavragani, Amaryllis, and Konstantinos Gkillas. “COVID-19 Predictability in the United States Using Google Trends Time Series.” *Scientific Reports*, vol. 10, no. 1, 27 Apr. 2020, mhttps://doi.org/10.1038/s41598-020-77275-9.

Purcell, Kristen, et al. “Search Engine Use 2012.*” Pew Research Center*, Washington, D.C. 9 Mar. 2012, https://www.pewresearch.org/internet/2012/03/09/search-engine-use-2012/.

Roozenbeek, Jon, et al. “Susceptibility to Misinformation about COVID-19 around the World.” *Royal Society Open Science*, vol. 7, no. 10, 14 Oct. 2020, p. 201199., https://doi.org/10.1098/rsos.201199.

Rovetta, Alessandro. “Reliability of Google Trends: Analysis of the Limits and Potential of Web Infoveillance during COVID-19 Pandemic and for Future Research.” *Frontiers in Research Metrics and Analytics*, vol. 6, 2021, https://doi.org/10.3389/frma.2021.670226.

Rufai, Sohaib R, and Catey Bunce. “World Leaders’ Usage of Twitter in Response to the COVID-19 Pandemic: A Content Analysis.” *Journal of Public Health*, vol. 42, no. 3, 2020, pp. 510–516., https://doi.org/10.1093/pubmed/fdaa049.

Schweinberger, Martin, et al. “Analysing Discourse around COVID-19 in the Australian Twittersphere: A Real-Time Corpus-Based Analysis.” *Big Data & Society*, vol. 8, no. 1, 30 May 2021, https://doi.org/10.1177/20539517211021437.

Shahi, Gautam Kishore, et al. “An Exploratory Study of Covid-19 Misinformation on Twitter.” *Online Social Networks and Media*, vol. 22, Mar. 2021, https://doi.org/10.1016/j.osnem.2020.100104.

Shearer, Elisa, and Katerina Eva Matsa. “News Use Across Social Media Platforms 2018.” Pew Research Center's Journalism Project, *Pew Research Center*, Washington, D.C. 10 Sep, 2018, https://www.pewresearch.org/journalism/2018/09/10/news-use-across-social-media- platforms-2018/.

Shin, Soo-Yong, et al. “High Correlation of Middle East Respiratory Syndrome Spread with Google Search and Twitter Trends in Korea.” *Scientific Reports*, vol. 6, no. 1, 6 Sept. 2016, https://doi.org/10.1038/srep32920.

Singh, Lisa, et al. “A First Look at COVID-19 Information and Misinformation Sharing on Twitter.” *ArXiv*. 31 Mar. 2020, https://arxiv.org/abs/2003.13907

Stephens-Davidowitz, Seth. *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us about Who We Really Are*, HarperColllins Publishers Inc., New York, 2017.

Tsai, Meng Hsiu, and Yingfeng Wang. “Analyzing Twitter Data to Evaluate People’s Attitudes towards Public Health Policies and Events in the Era of Covid-19.” *International Journal of Environmental Research and Public Health*, vol. 18, no. 12, 2021, p. 6272., https://doi.org/10.3390/ijerph18126272.

Turk, Philip J., et al. “A Predictive Internet-Based Model for Covid-19 Hospitalization Census.” *Scientific Reports*, vol. 11, no. 1, 3 Mar. 2021, https://doi.org/10.1038/s41598-021-84091-2.

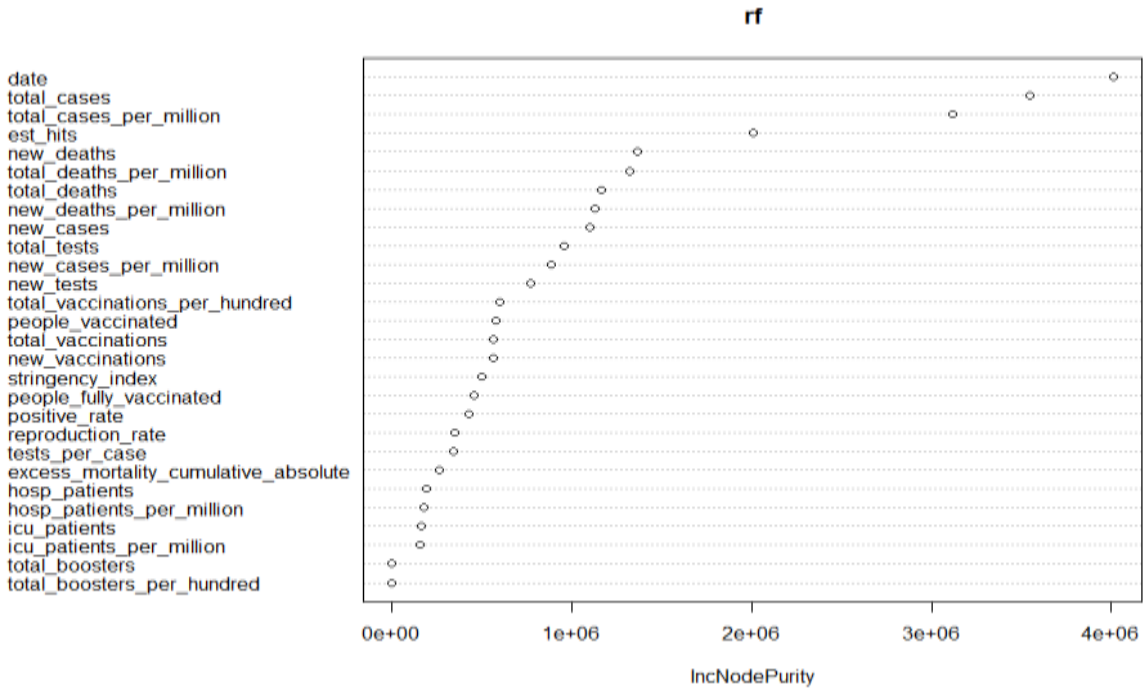
Yousefinaghani, Samira, Dara, Rozita, Mubareka, Samira and Sharif, Shyan. “Prediction of Covid-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada.” Frontiers in Public Health, 16 April 2021, https://doi.org/10.3389/fpubh.2021.656635.

Zepecki, Anne, et al. “Using Application Programming Interfaces to Access Google Data for Health Research: Protocol for a Methodological Framework.” *JMIR Research Protocols*, vol. 9, no. 7, 2020, https://doi.org/10.2196/16543.

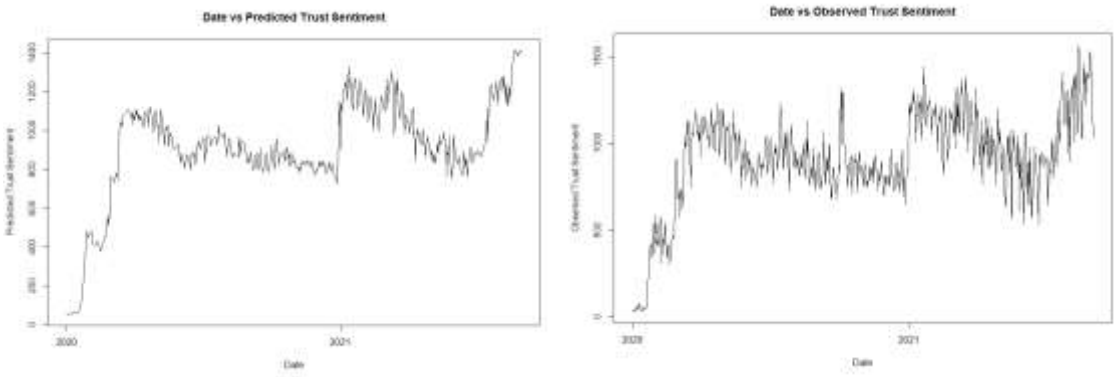
Zhang, Xiongwei, et al. “Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System.” *Complexity*, vol. 2020, 2020, pp. 1–10., https://doi.org/10.1155/2020/6688912.

Appendices

Trust Sentiment Random Forest

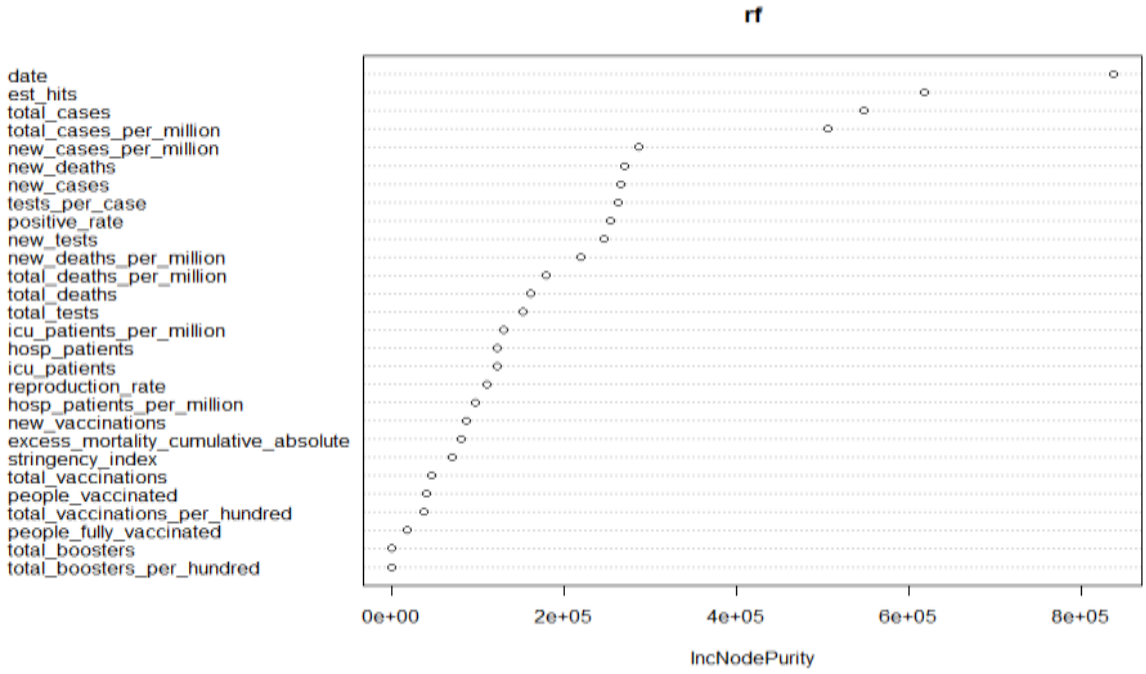


Appendix 1: Variable important plot for trust sentiment random forest using node purity

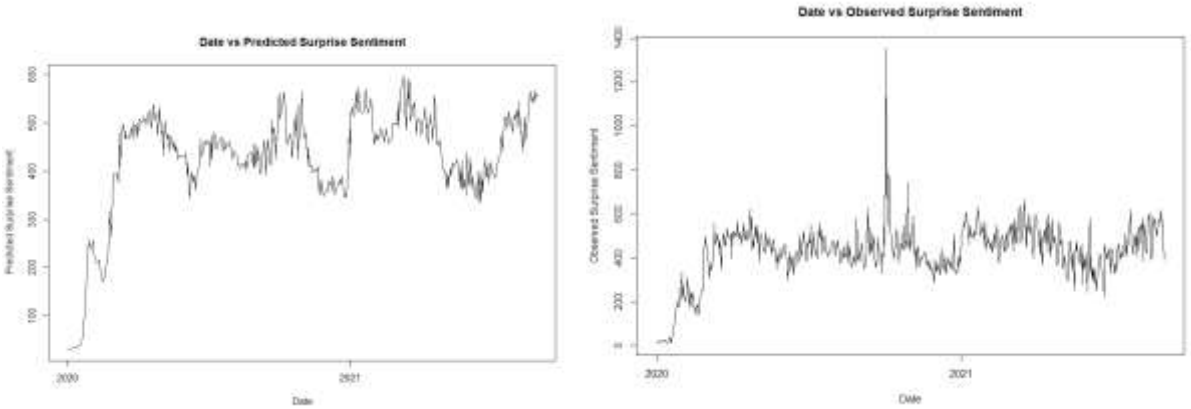


Appendix 2: Observed (left) and predicted (right) trust sentiment vs time

Surprise Sentiment Random Forest

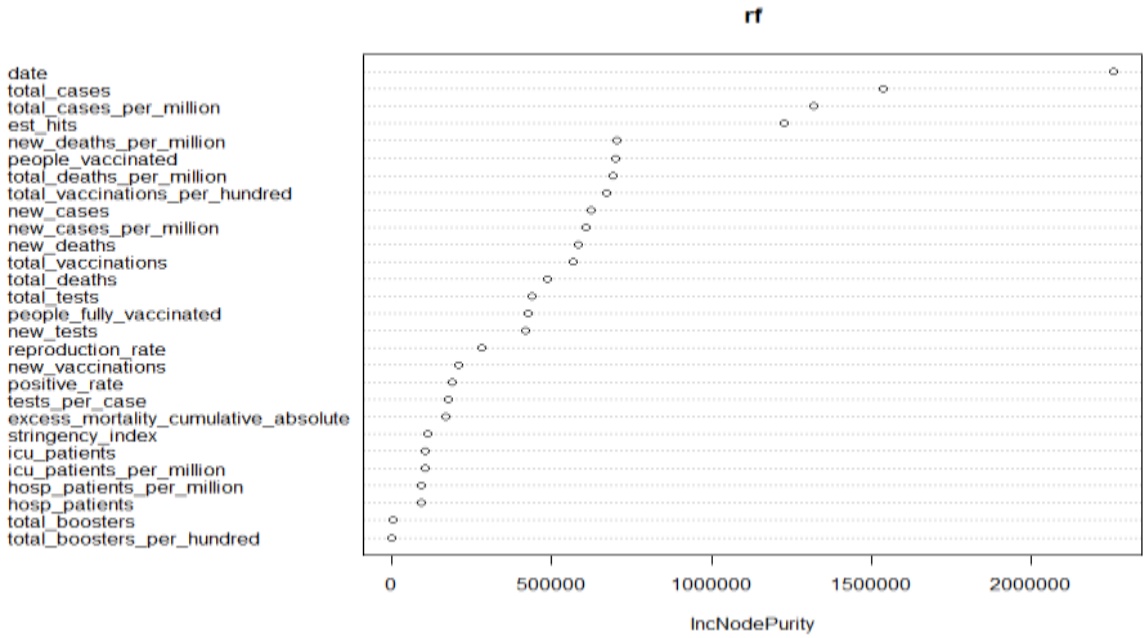


Appendix 3: Variable important plot for surprise sentiment random forest using node purity

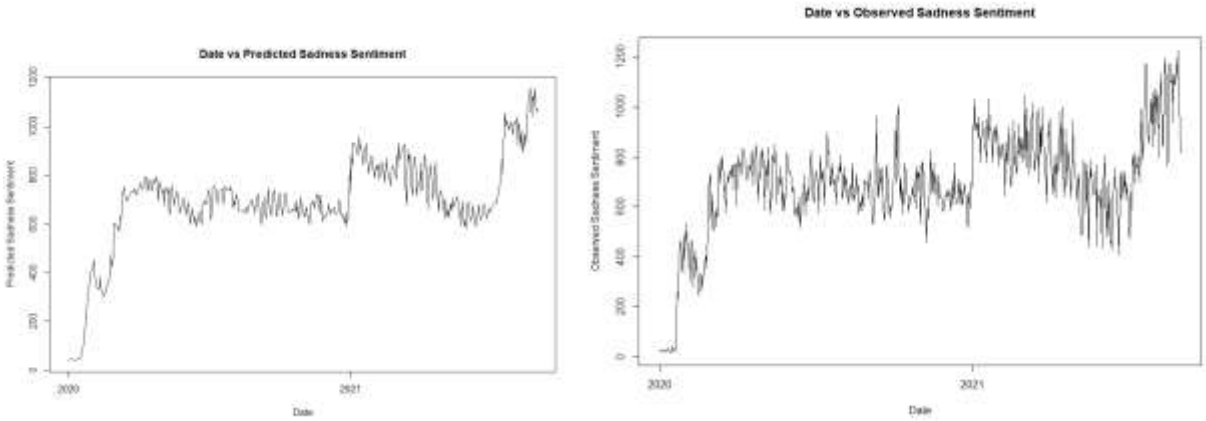


Appendix 4: Observed (left) and predicted (right) surprise sentiment vs time

Sadness Sentiment Random Forest

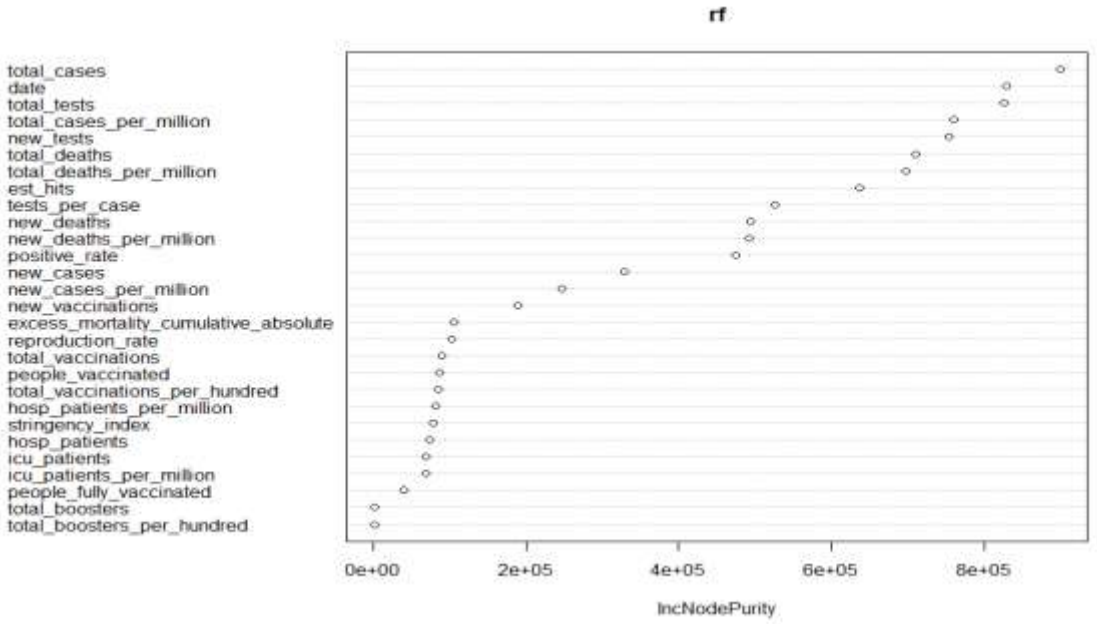


Appendix 5: Variable important plot for sadness sentiment random forest using node purity

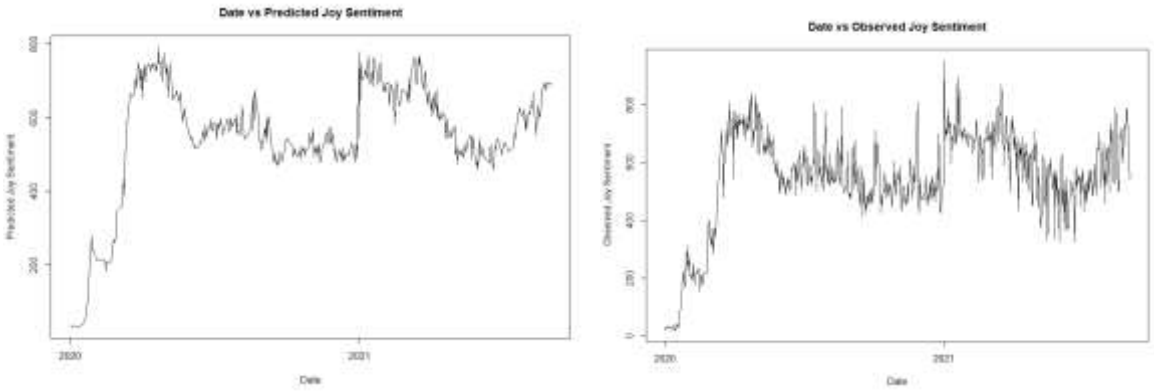


Appendix 6: Observed (left) and predicted (right) sadness sentiment vs time

Joy Sentiment Random Forest

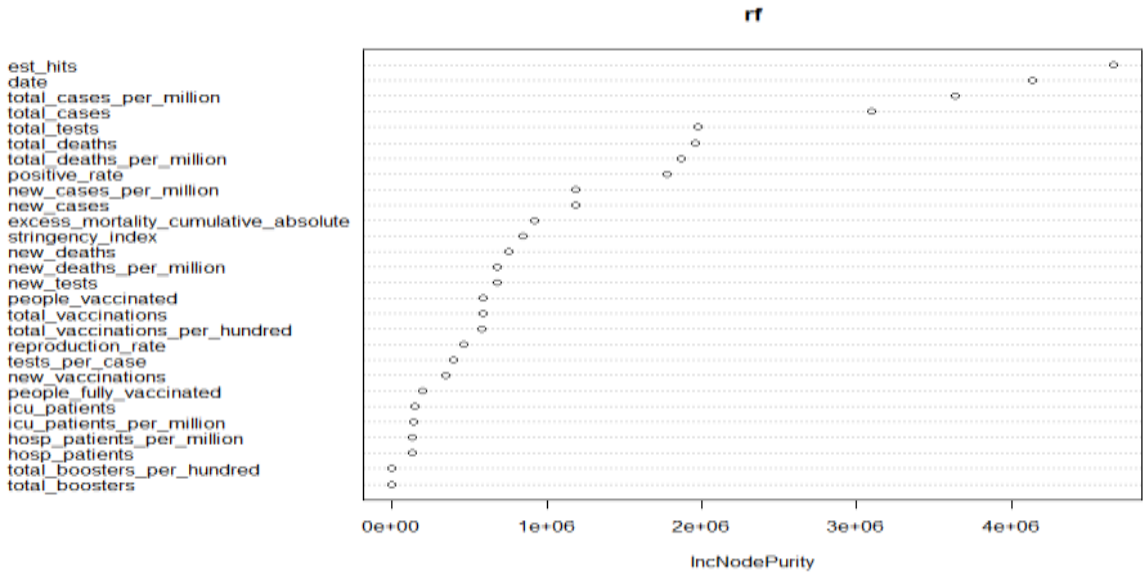


Appendix 7: Variable important plot for joy sentiment random forest using node purity

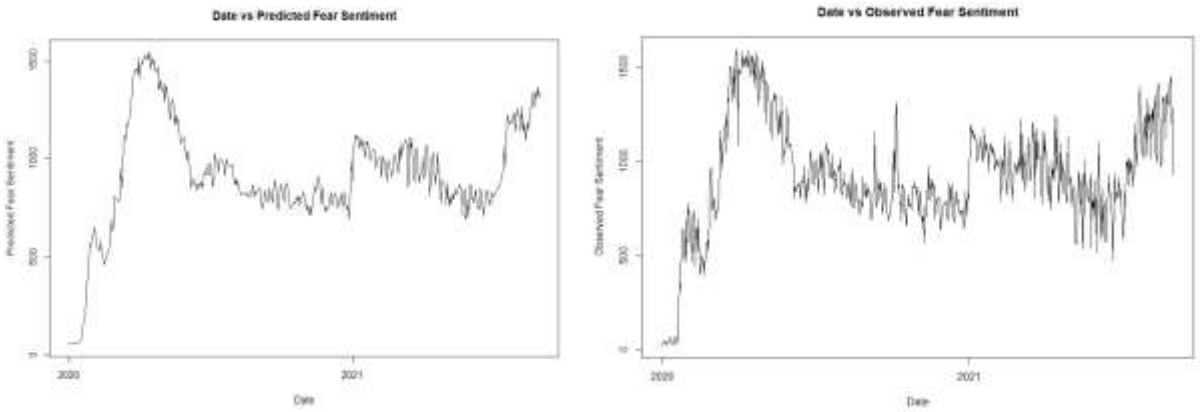


Appendix 8: Observed (left) and predicted (right) joy sentiment vs time

Fear Sentiment Random Forest

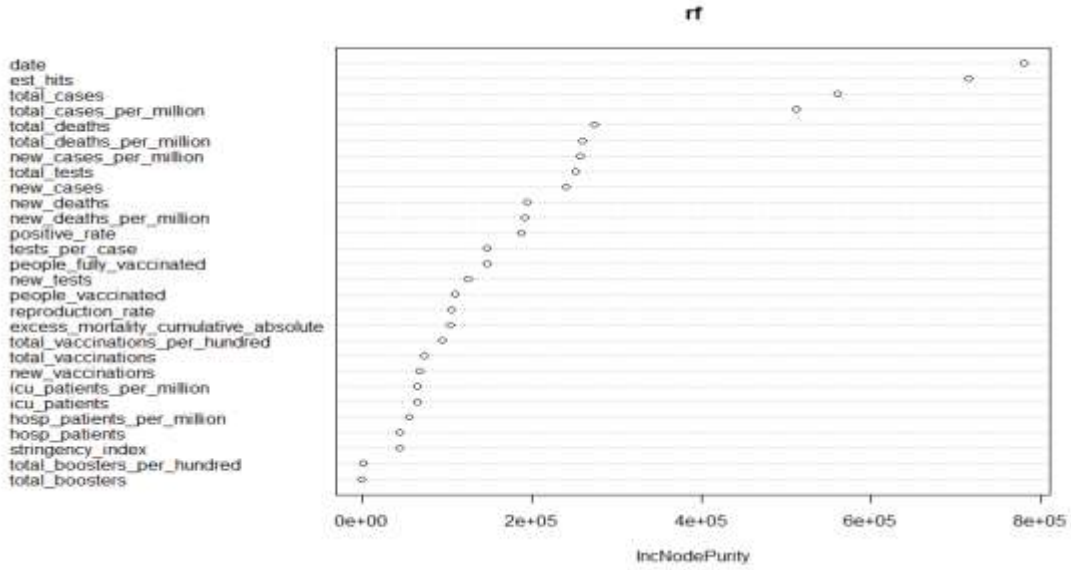


Appendix 9: Variable important plot for fear sentiment random forest using node purity

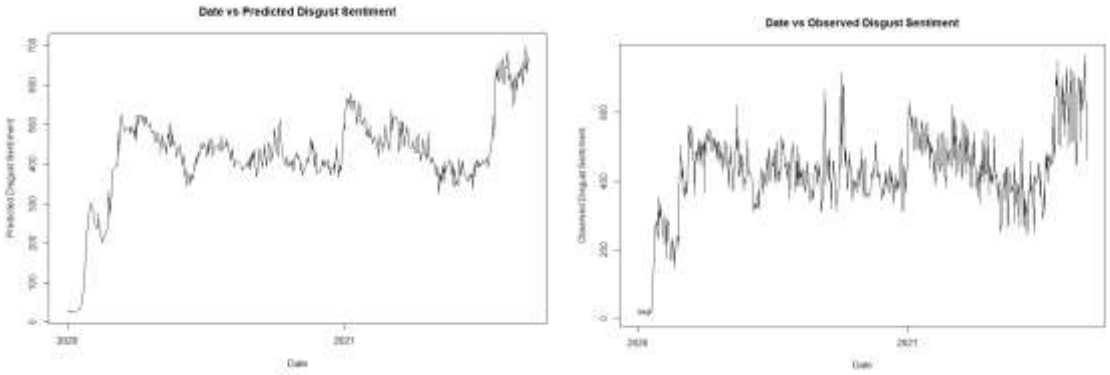


Appendix 10: Observed (left) and predicted (right) fear sentiment vs time

Disgust Sentiment Random Forest

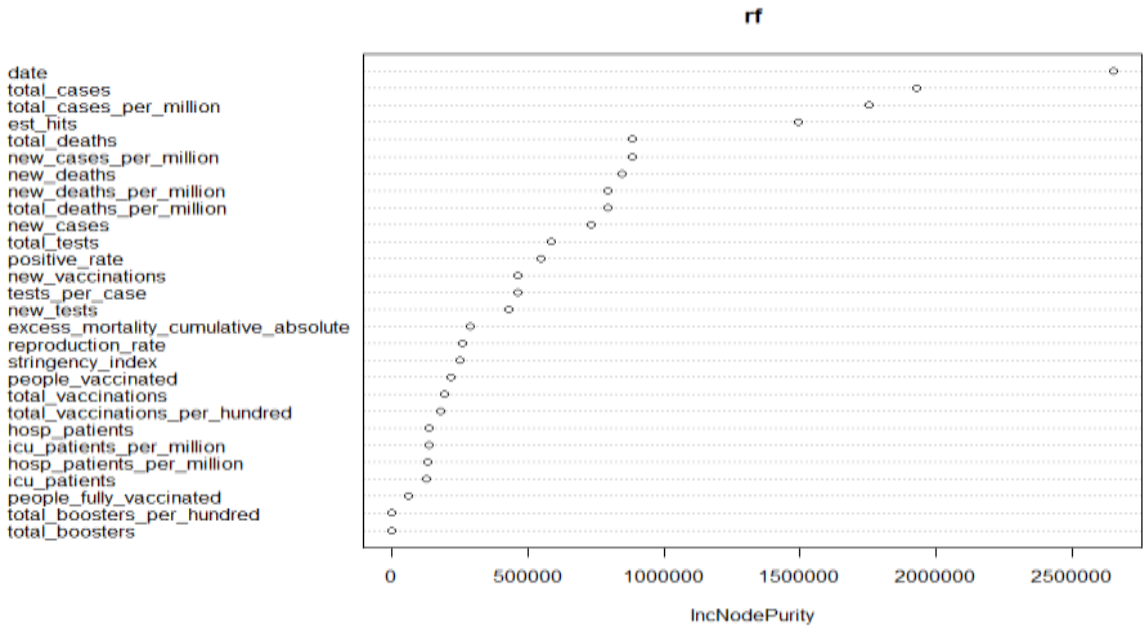


Appendix 11: Variable important plot for disgust sentiment random forest using node purity

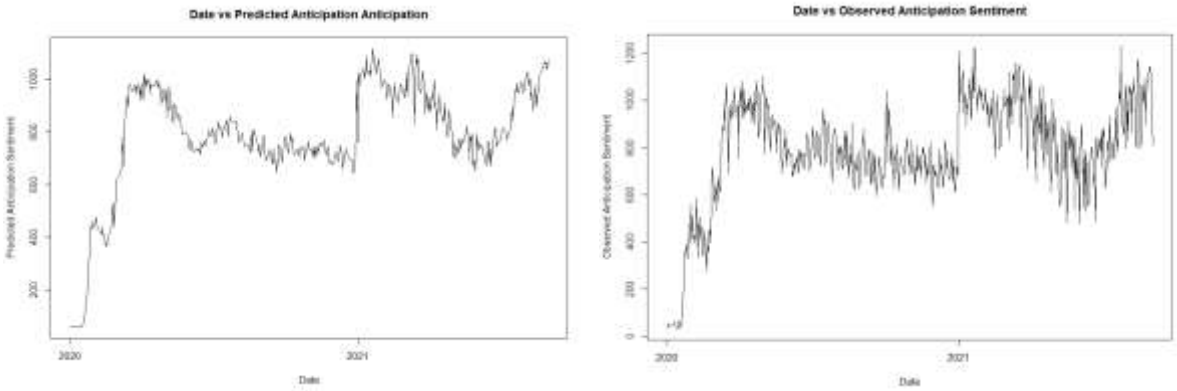


Appendix 12: Observed (left) and predicted (right) disgust sentiment vs time

Anticipation Sentiment Random Forest

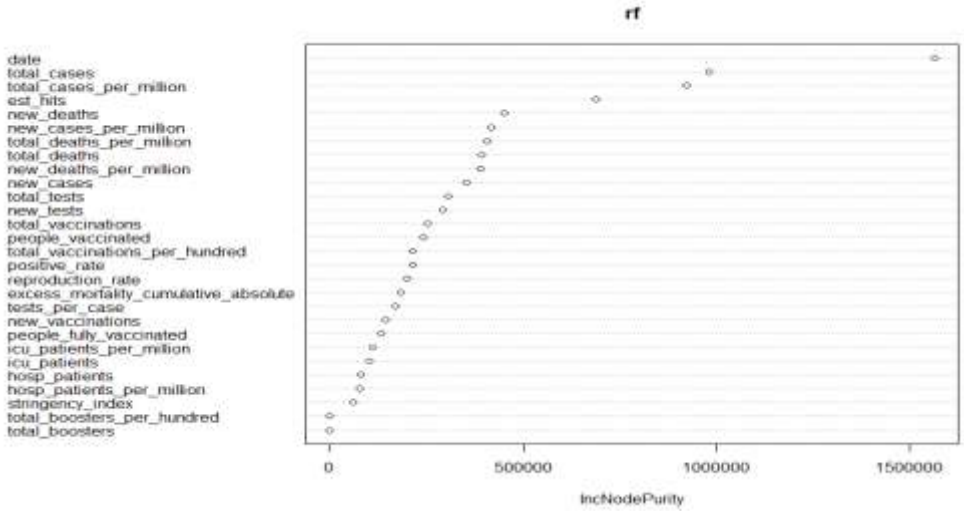


Appendix 13: Variable important plot for anticipation sentiment random forest using node purity

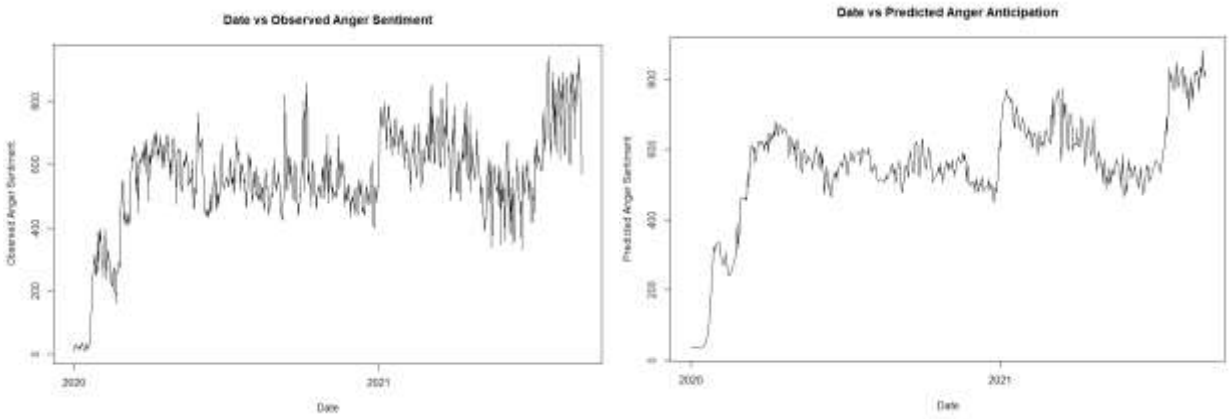


Appendix 14: Observed (left) and predicted (right) anticipation sentiment vs time

Anger Sentiment Random Forest



Appendix 15: Variable important plot for anger sentiment random forest using node purity



Appendix 16: Observed (left) and predicted (right) anger sentiment vs time