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 Havis Rathke MS[1], Han Yu PhD[1] and Hong Huang PhD[2]

[1]Department of Applied Statistics and Research Methods, University of Northern Colorado, Greeley, Colorado, USA and [2]School of Information, University of South Florida, Tampa, Florida, USA

Abstract

Objective: The rapid onset of coronavirus disease 2019 (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 have become valuable for monitoring and understanding public sentimental preparedness and response to an event within our society.

Methods: 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 8 positive and negative sentiments and emotions. Machine learning algorithms were used to implement the opinion mining how Twitter sentiment was related to Google Trends interest with historical COVID-19 public health data.

Results: The sentiment analysis went beyond polarity to detect specific feelings and emotions during the pandemic.

Conclusions: 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.

SDMPH

SOCIETY FOR DISASTER MEDICINE & PUBLIC HEALTH

The novel coronavirus disease 2019, 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 (ie, 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."[1] 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."[2] 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 2-fold increase in ages 50 and older who report using at least 1 app.[3] 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 approximately 1 in 5 say that it is their primary source of news.[4,5] 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.[5] Rufai and Bunce[6] 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.[7] assert that more than 4 in 5 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.[8] 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[4] 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."[2] 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 (eg, population surveys)."[9] A total of 83% of Americans use Google as their main search engine, making Google the most popular search engine in the United States.[10] 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 a machine learning approach. With the rapid spread of misinformation during the pandemic, it remains to be known 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.[11] COVID-19 discourse has recently been examined on Twitter by means of frequency analysis of likes, comments and retweets, word-cloud mapping, stance detection, sentiment analysis, and network modeling.[6,12–14] 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.[13,15–17] Schweinberger et al.[15] chose to model topics and sub-topics across different phases of the pandemic. Singh et al.[18] demonstrated that Twitter conversations may be used to predict the spread and outbreak of COVID-19. Hu et al.[13] and Hussain et al.[16] 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.[19] 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.[20] found high correlations between the number of confirmed MERS cases and Twitter sentiment and Google interest. For the COVID-19 pandemic, Diaz and Henriquez[21] compared Twitter sentiment and Google interest with fluctuations in the stock market and number of people under lockdown. Mavragani and Gkillas[22] 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[23] used Twitter and Google Trends to analyze search behaviors and predict new COVID-19 cases.

Zhang et al.,[24] furthermore, demonstrated that machine learning, specifically a unigram random forest (RF) model, is a powerful tool to predict coronavirus sentiment. RF regression models tend to outperform classical approaches in analyzing highly non-linear and complex relationships.[25] Cornelius et al.[26] used RFs to predict COVID-19 patient mortality. Iwendi et al.[27] used RF models to predict severity of COVID-19 cases using patient geographical, travel, health, and demographic data. RFs 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 (ie, cases, deaths, vaccinations, positive tests, hospitalizations, school closures, travel bans, etc.) and Google Trends data in determining social media sentiment. RFs 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.[28–30] There may be evidence that negative events are more contagious than positive events.[28] On the other hand, certain key topics relating to the pandemic may be perceived more positively than expected. Yousefinaghani et al.[31] 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.[32] 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.[33] 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 8 twitter emotion types and their dynamics over time using the data from January 1, 2020, to September 1, 2021, in the United States. 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 daily 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.[34] 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 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.[15] and Hu et al.[13] Future studies may first do a relevant topic analysis, then pull relevant tweets for a more representative sample. A unigram (1-word) method was chosen because of its optimal use in RF models.[24] A total of 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 3 locations throughout each day as outlined by Kim et al.[35] 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.[36] 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 were 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 mo, 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 were 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 intensive care unit (ICU), tests and positivity, the reproduction rate of the virus, policy responses, and other variables of interest. Missing data were substituted with estimated values from near neighbors as outlined by Kang.[37] (2013). New cases and new deaths over time are visualized in Figures 2 and 3, respectively.

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.

Methods

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 8 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 RF models. Twitter sentiment counts, Google Trends estimated interest, and historical COVID-19 data were aggregated by day, and 10 RF 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 RF 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 RFs.

Random Forests

RFs are a substantial modification of bootstrap aggregation (bagging). Bagging is a variance-reduction technique for an estimated predictive function, formed by building a large collection of de-correlated trees with each generated tree being identically distributed, then averaging the resulting trees.[38] Trees are ideal candidates for sentiment analysis because 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.


Algorithm of RF

1. For b = 1 to B:
   (I) Draw a bootstrap sample Z^* of size N from the training data.
   (II) Grow an RF 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 2 daughter nodes.
2. Output the ensemble of trees $\{T_b\}_{b=1}^{B}$.
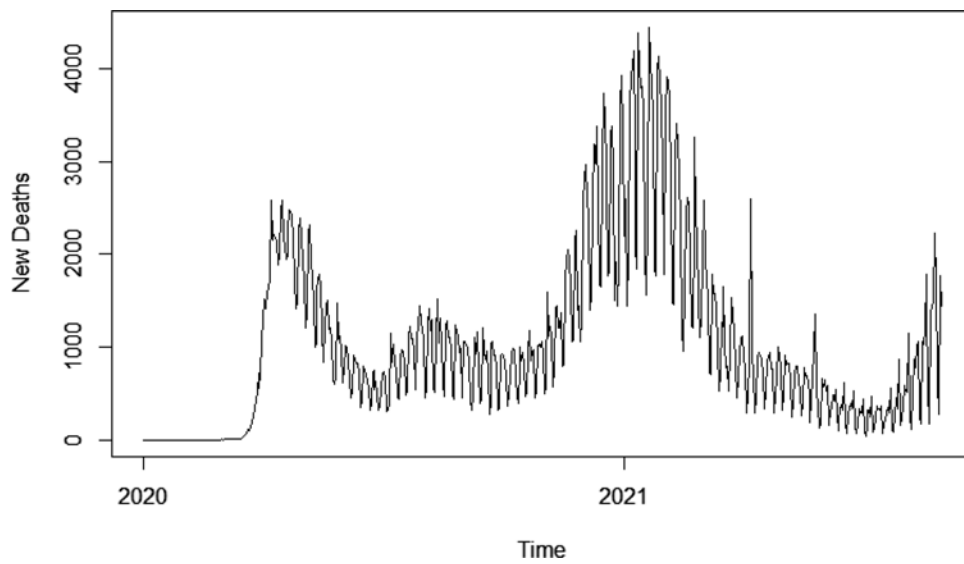3. $\hat{f}_{rf}^{B}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$.

After Bth recursion, tree sequences $\{T(x; \theta_b(Z)\}_{b=1}^{B}$ are grown, the RF predictor at a single target point x is
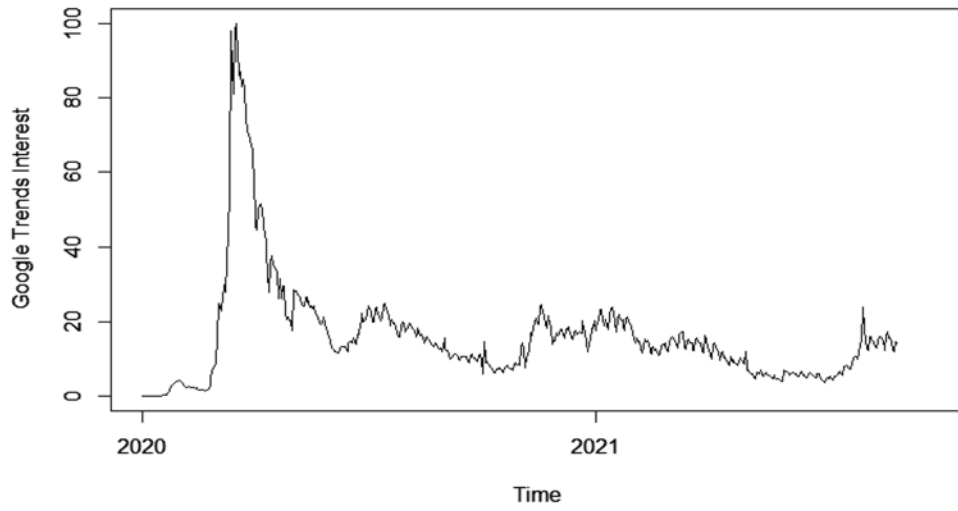
Figure 1. Google Trends interest over time in the United States.
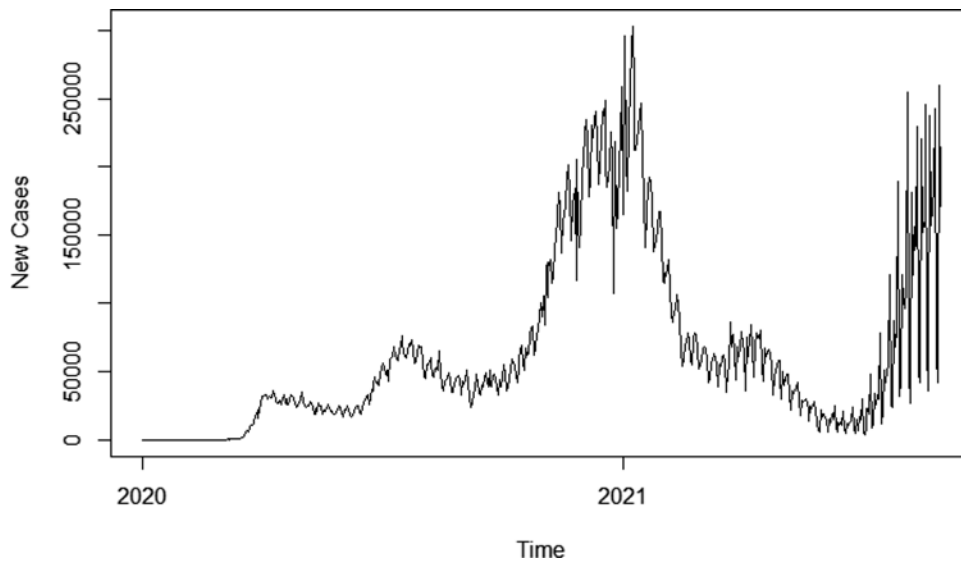
Figure 2. New COVID-19 cases over time in the United States.

Figure 3. New COVID-19 deaths over time in the United States.

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

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x;\ \theta_b(Z)),$$

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

Increasing B does not cause the RF to overfit as

$$\hat{f}_{rf}(x) = E_{\theta|Z} T(x;\theta(Z)) = \lim_{B\to\infty} \hat{f}_{rf}^{B}(x)$$

with an average over B realizations of $\theta(Z)$ and the distribution of $\theta(Z)$ is conditional on the training data Z. Using full-grow trees results 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 $\rho(x)$ is the conditional sampling correlation between any pair of trees used in the averaging,

$$\rho(x) = cor(T(x;\theta_1(Z), T(x;\theta_2(Z));$$

where $\theta_1(Z)$ and $\theta_2(Z)$ are a randomly drawn pair of RF trees grown to the randomly sampled Z. $\sigma^2(x)$ is the sampling variance of any single randomly drawn tree, $\sigma^2(x)=Var(T(x;\theta(Z)).$
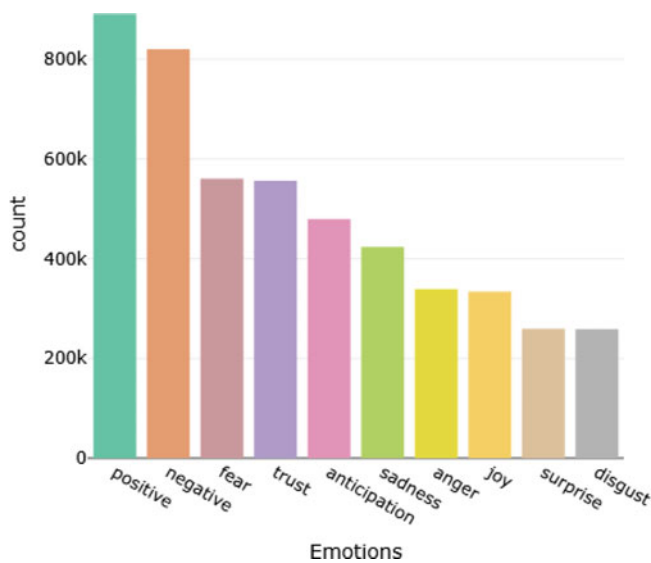
Figure 4. Emotion type for COVID-19 tweets.

Then

$$\mathrm{Var}(\hat{f}_{rf}\,(x)) = \rho(x)\sigma^2(x).$$

The conditional covariance of a pair of trees 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 RFs is very similar to boosting, and they are simpler to train and tune. Hastie et al.[39] made grand claims that RFs 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.

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.

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 RF (420 observations), and Figure 8 used the predicted values of each RF 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.
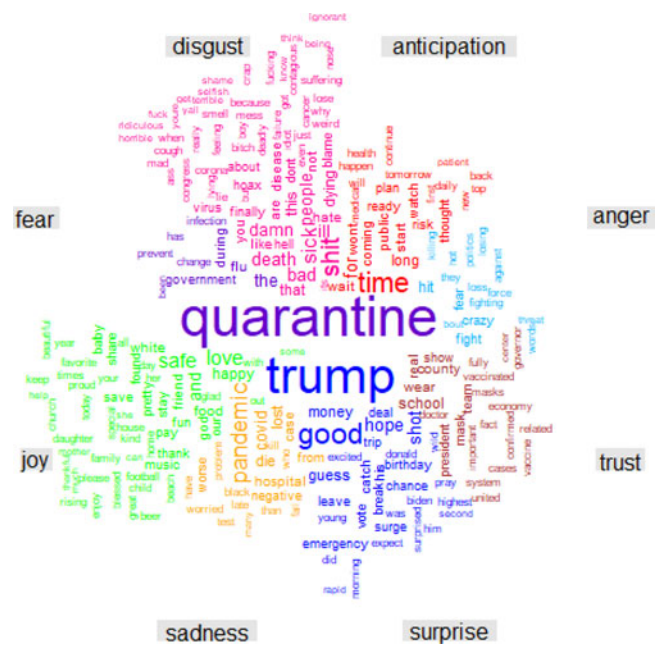

Figure 5. Word Cloud for emotional terms in the United States.

Visually, it appears that the predictive models performed quite well, matching with the actual data. The MAPE and the reported percentage of variation explained, quantified how well the RF 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 2 to 3 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.

The important variables for each RF model are now detailed for each sentiment type with plots of observed and predicted sentiment provided for reference.

Results

Positive Sentiment Random Forest

The variable importance plot for positive sentiment RF are shown in Figure 9. As a proof of concept, minimal depth and frequent interactions are also plotted to compare important variables decided by node purity and are shown in Figures 10 and 11. With cross-validation of the mean of minimal depth distribution and interaction (Figures 10 and 11), among all the variables, "date", "total_cases_per_million", "total_cases", and "est_hits" are important ones for positive sentiment RF.

Node purity and minimal depth provided similar results for deciding important variables. 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 start of the pandemic, then was stable later; another wave was observed starting in 2021.
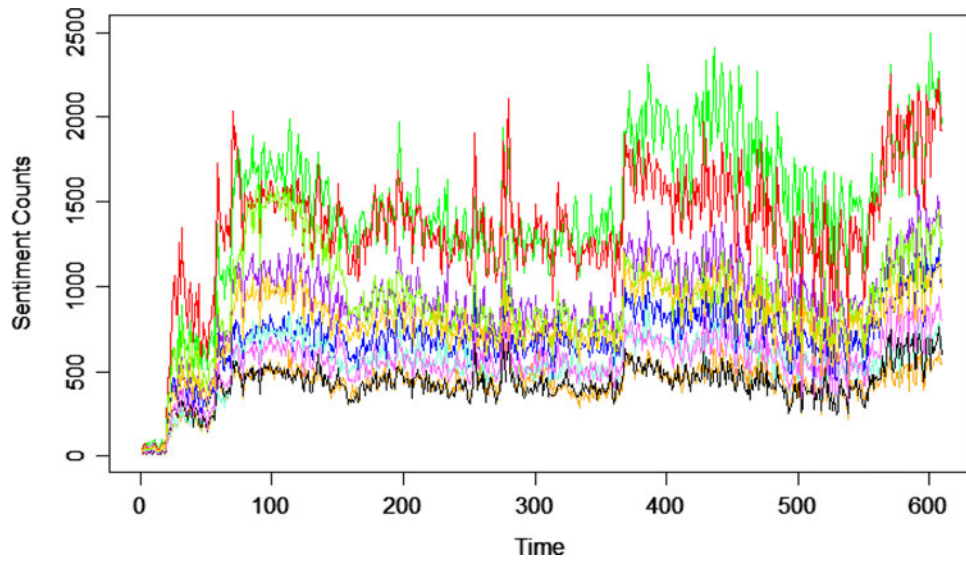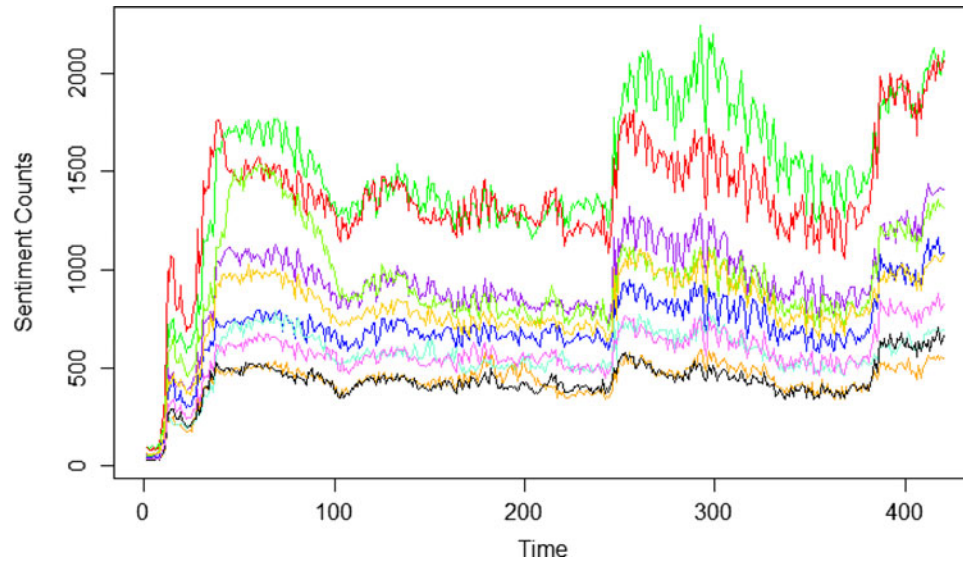
Figure 6. Trajectories of observed sentiment counts over time in the United States.
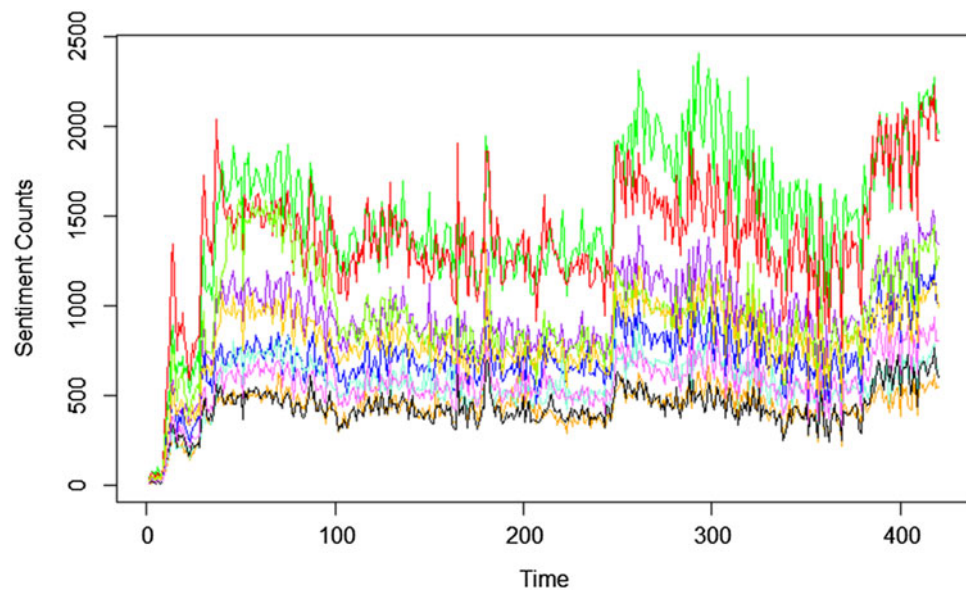

Figure 7. Trajectories of training data from observed sentiment counts over time in the United States.

Table 2. RF 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%

Figure 8. Trajectories of predicted sentiment counts over time in the United States.
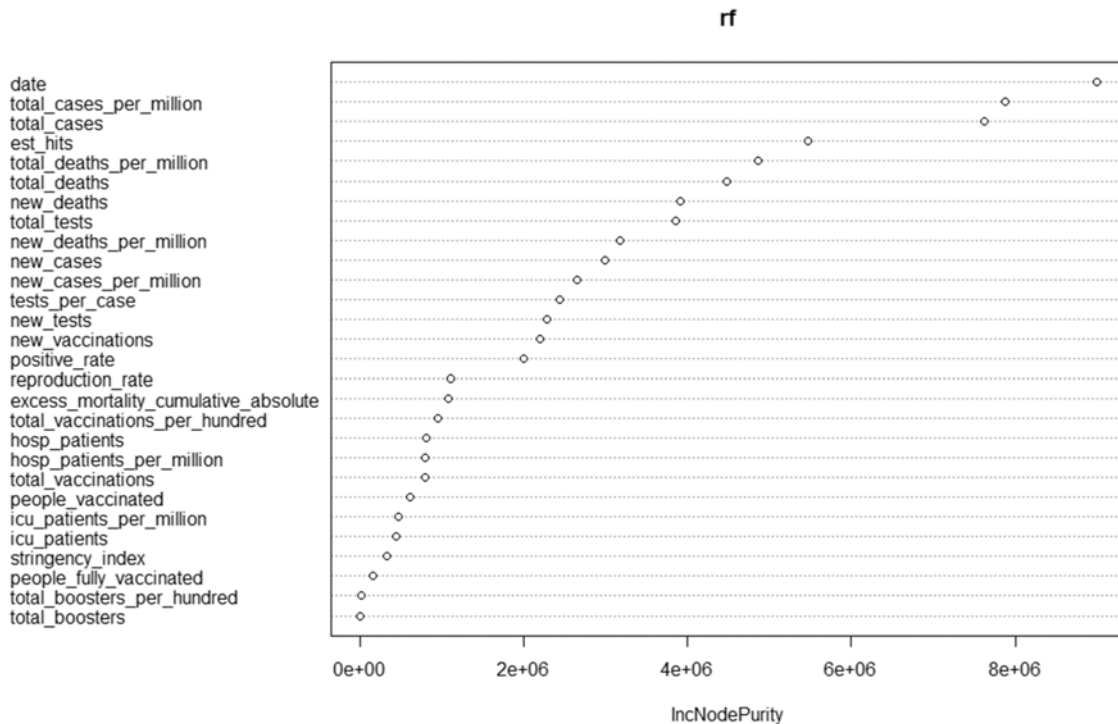
Figure 9. A variable importance plot for positive sentiment RF using node purity.

Negative Sentiment Random Forest

The variable importance plot for negative sentiment RF are shown in Figure 13. "est_hits", "date", "total_cases", and "total_cases_per_million" are the important variables for negative sentiment RF. Notably, Google Trends interest appears to be the most important variable for prediction.

The observed and predicted negative sentiment over time are shown in Figure 14. The negative sentiments increased at the beginning of the COVID-19, with fluctuation over 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 COVID-19 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). Of interest, fear sentiment appears in the first wave at the start of COVID-19, then falls noticeably, and then returns with a spike at the end of 2021, but at a lower level than the initial jump (Figure 17).
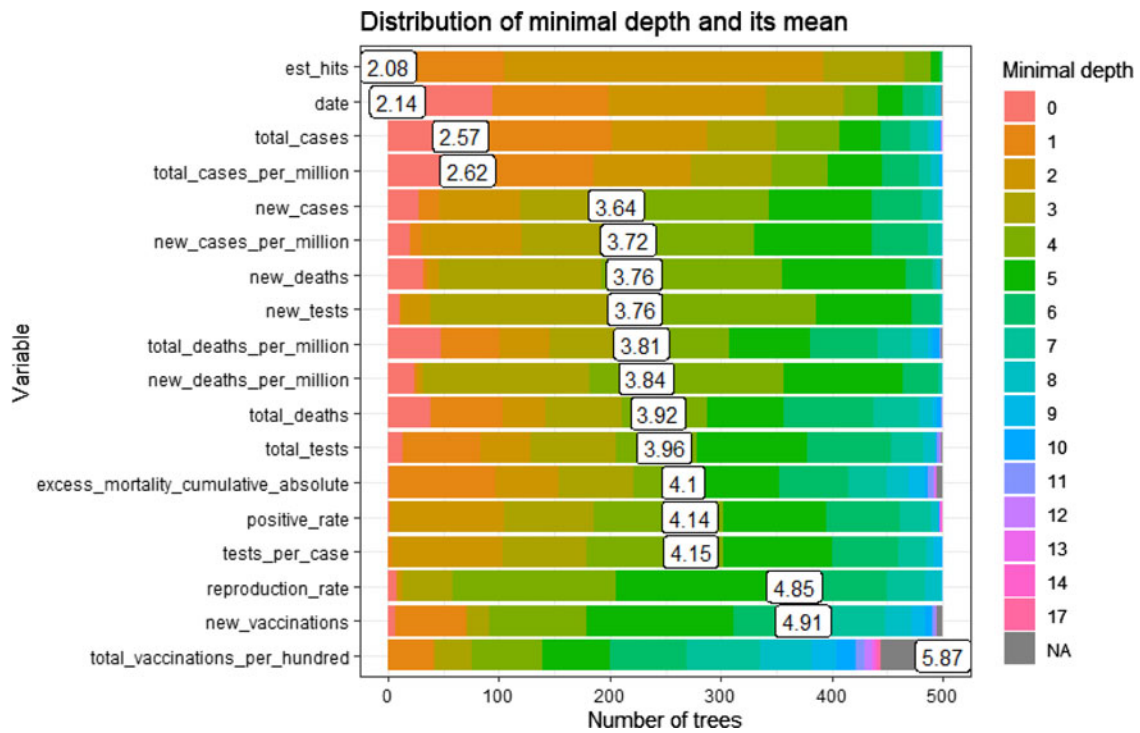
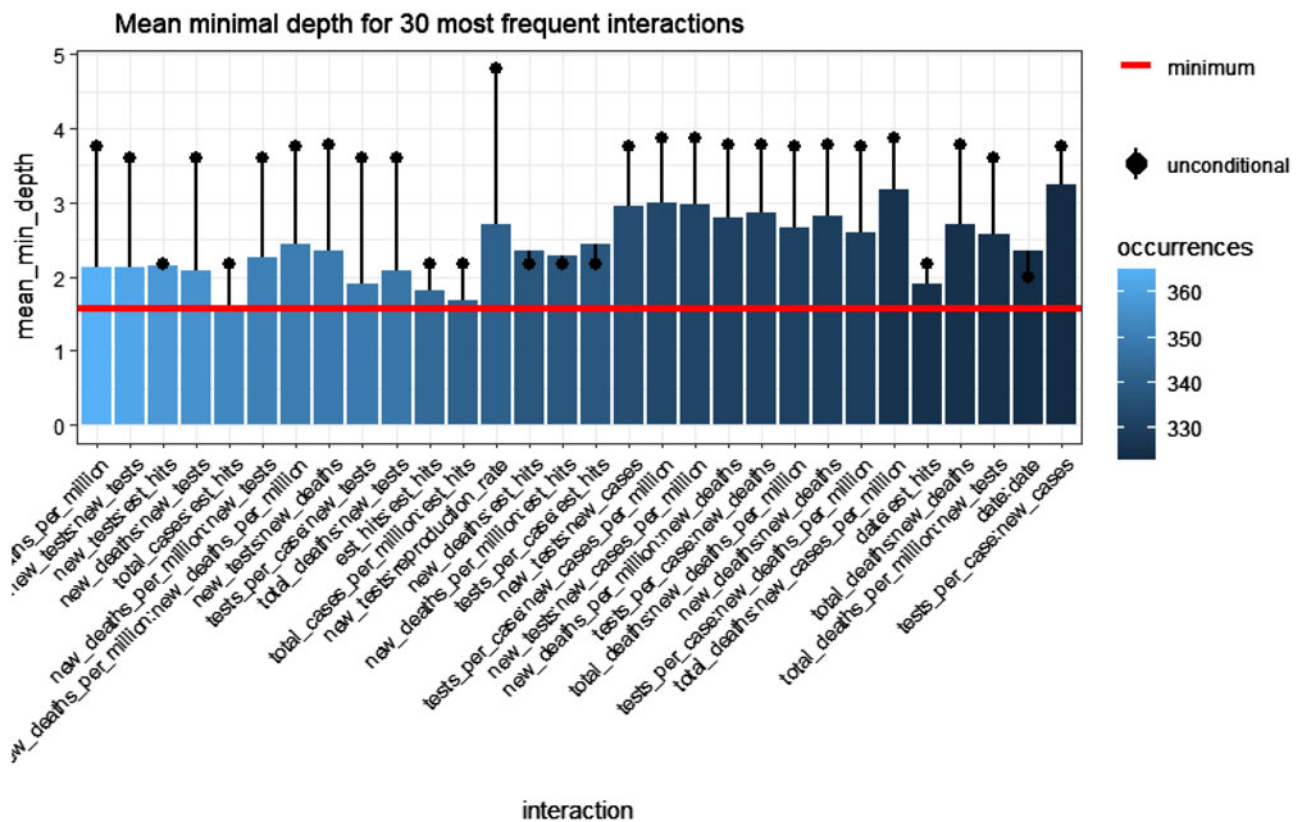**Figure 10.** A variable importance plot for positive sentiment RF using distribution of minimal depth and its mean.

Figure 11. A variable importance plot for positive sentiment RF using interaction.
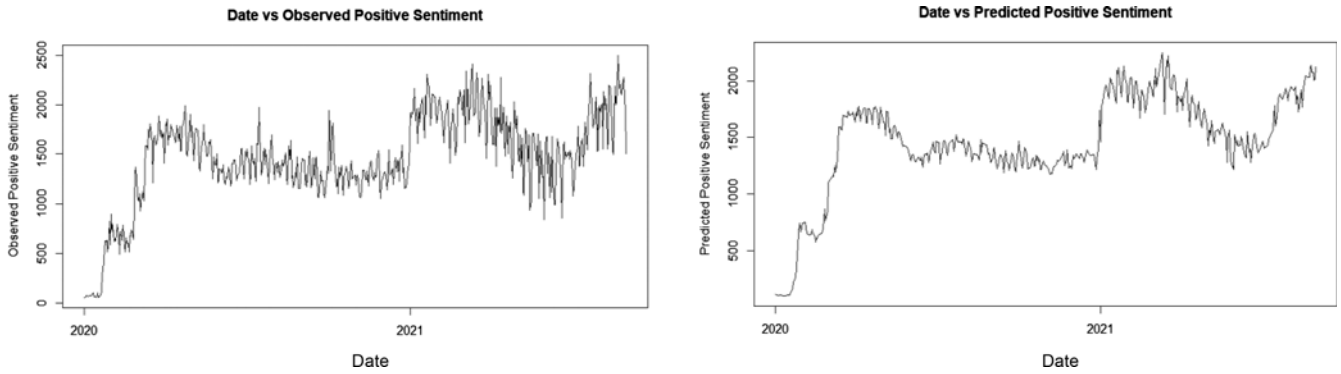
Figure 12. Observed (left) and predicted (right) positive sentiment vs time.
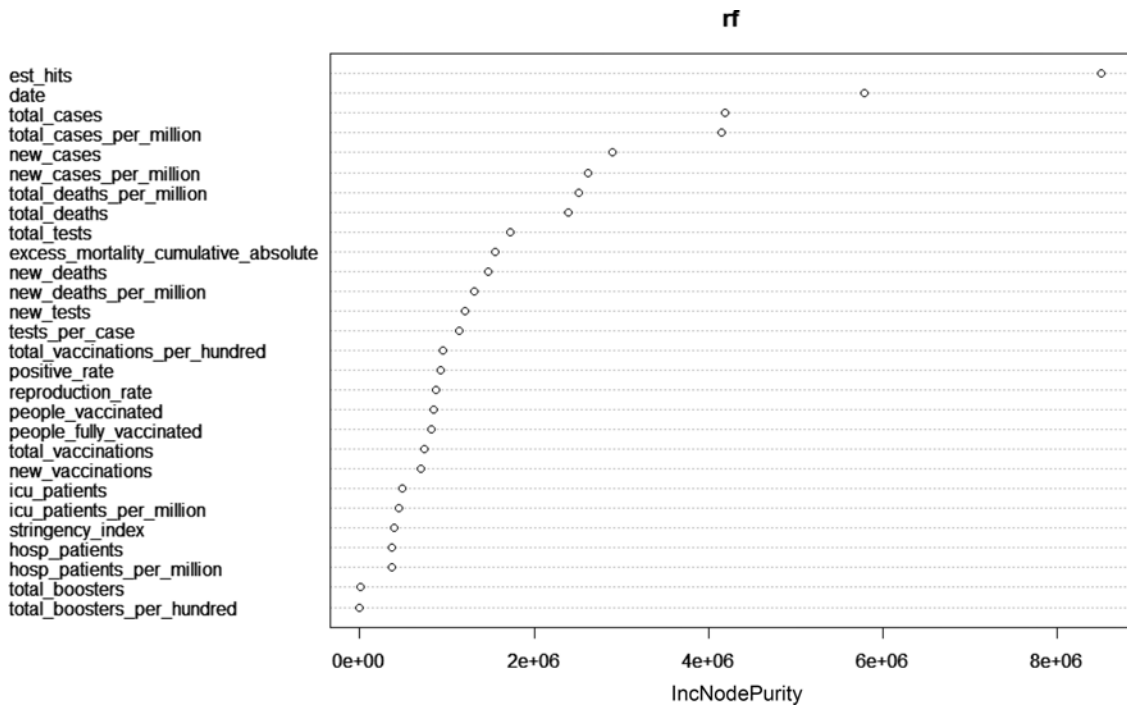

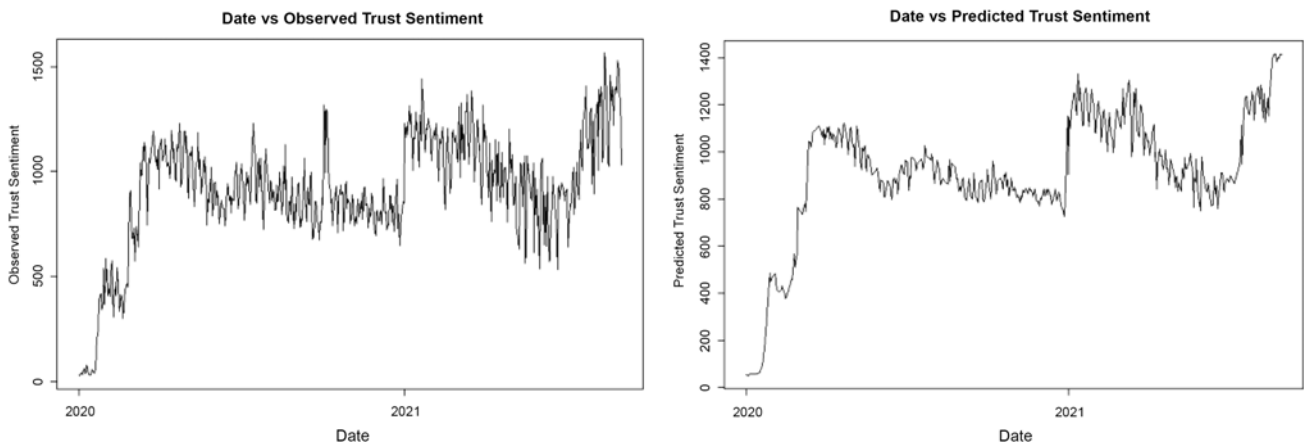**Figure 13.** Variable importance plot for negative sentiment RF using node purity.
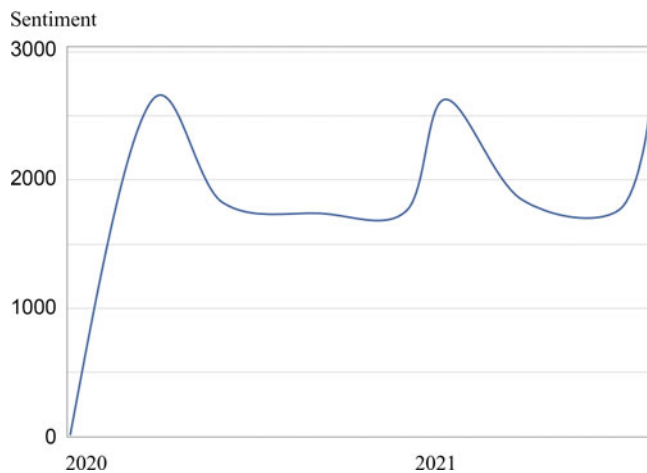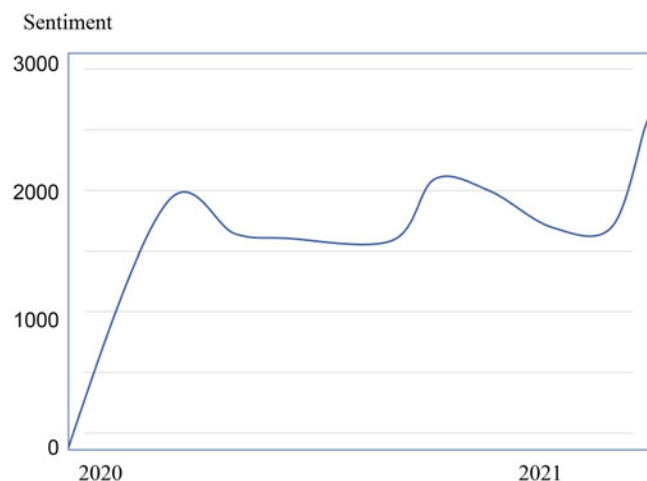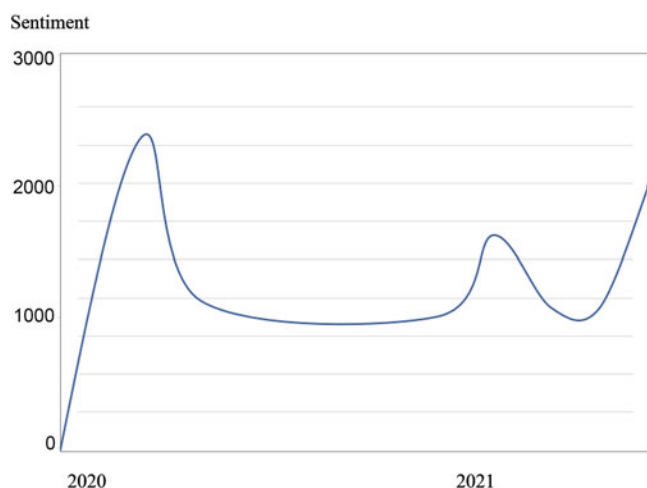

Figure 14. Observed (left) and predicted (right) negative sentiment vs time.

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.,[13] Hussain et al.,[16] and Ahmed et al.[19]

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,[22] Turk et al.,[40] and Alshahrani and Babour.[23]

RF 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 COVID-19 might be linked to the recovery progress, vaccine development, new hopes of technologies development, and resilience.[33]

Anger, disgust, and sadness sentiments increased during the pandemic, indicating that people in the United States emotionally were not expecting such a long duration of the pandemic. Fear sentiment shows a big wave at the beginning of COVID-19 since 2020, later on drops gradually, then has 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, so like the positive sentiment trend, it demonstrated a flat and wavy behavior, reflecting hope at the beginning of 2020 when COVID-19 starts, and at the beginning of 2021. Anticipation, surprise, and negative sentiments showed a series of fluctuation waves. This appears to indicate that people were invested in analysis and information seeking behaviors as evidenced 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 focused on text-based emotion analysis 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.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/dmp.2023.101

Competing interests. The authors declare that there is no conflicts of interest.

References

1. **Athique A.** Extraordinary issue: coronavirus, crisis and communication. *Media Int Aust.* 2020;177(1):3-11. doi: 10.1177/1329878X20960300
2. **Stephens-Davidowitz S.** *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us about Who We Really Are.* HarperColllins; 2017.
3. **Pew Research Center.** Demographics of social media users and adoption in the United States." 2021. Accessed June 25, 2023. https://www.pewresearch.org/internet/fact-sheet/social-media/?menuItem=b14b718d-7ab6-46f4-b447-0abd510f4180
4. **Infield T.** *Americans who get news mainly on social media are less knowledgeable and less engaged.* The Pew Charitable Trusts, Washington, DC. 2020. Accessed June 25, 2023. https://www.pewtrusts.org/en/trust/archive/fall-2020/americans-who-get-news-mainly-on-social-media-are-less-knowledgeable-and-less-engaged
5. **Shearer E, Matsa KE.** News use across social media platforms 2018. *Pew Research Center's Journalism Project.* Pew Research Center, Washington, DC. 2018. Accessed June 25, 2023. https://www.pewresearch.org/journalism/2018/09/10/news-use-across-social-media-platforms-2018/
6. **Rufai SR, Bunce C.** World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. *J Public Health (Oxf).* 2020; 42(3):510-516. doi: 10.1093/pubmed/fdaa049
7. **Shahi GK, Dirkson A, Majchrak TA.** An exploratory study of Covid-19 misinformation on Twitter. *Online Soc Netw Media.* 2021;22:100104. doi:10.1016/j.osnem.2020.100104
8. **Roozenbeek J, Schneider CR, Dryhurst S, *et al.*** Susceptibility to misinformation about COVID-19 around the World. *R Soc Open Sci.* 2020;7(10):201199. doi: 10.1098/rsos.201199
9. **Rovetta A.** 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.* 2021;6:670226. doi: 10.3389/frma.2021.670226
10. **Purcell K, Brenner J, Rainie L.** *Search engine use 2012.* Pew Research Center, Washington, DC. 2012. Accessed June 25, 2023. https://www.pewresearch.org/internet/2012/03/09/search-engine-use-2012/
11. **Bossetta M.** The digital architectures of social media: comparing politicalcampaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. Election. *Journal Mass Commun Q.* 2018;95(2):471-496. doi: 10.1177/1077699018763307
12. **Tsai MH, Wang Y.** Analyzing Twitter data to evaluate people's attitudes towards public health policies and events in the era of Covid-19. *Int J Environ Res Public Health.* 2021;18(12):6272. doi: 10.3390/ijerph18126272
13. **Hu T, Wang S, Luo W, *et al.*** Revealing public opinion towards Covid-19 vaccines with Twitter data in the United States: a spatiotemporal perspective. *J Med Internet Res.* 2021;23(9):e30854. https://doi.org/10.2196/30854
14. **Fuentes A, Peterson JV.** Social media and public perception as core aspect of public health: the cautionary case of @realdonaldtrump and Covid-19. *PLoS One.* 2021;16(5):e0251179. doi: 10.1371/journal.pone.0251179
15. **Schweinberger M, Haugh M, Hames S.** Analysing discourse around COVID-19 in the Australian Twittersphere: a real-time corpus-based analysis. *Big Data Soc.* 2021. doi: 10.1177/20539517211021437
16. **Hussain A, Tahir A, Hussain Z, *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. *J Medl Internet Res.* 2021;23(4):e26627. doi: 10.2196/26627
17. **Lyu JC, Luli GK.** Understanding the public discussion about the Centers for Disease Control and Prevention during the COVID-19 pandemic using Twitter data:text mining analysis study. *J Med Internet Res.* 2021;23(2): e25108. doi: 10.2196/25108
18. **Singh L, Bansal S, Bode L, *et al.*** A first look at COVID-19 information and misinformation sharing on Twitter. *ArXiv.* 2020;arXiv:2003.13907v1
19. **Ahmed ME, Rabin RI, Chowdhury FN.** Covid-19: social media sentiment analysis on reopening. *ArXiv.* arxiv:2006.00804.
20. **Shin S-Y, Seo D-W, An J, *et al.*** High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. *Sci Rep.* 2016;6: 32920. doi: 10.1038/srep32920
21. **Díaz F, Henríquez PA.** Social sentiment segregation: evidence from Twitter and Google trends in Chile during the COVID-19 dynamic quarantine strategy. *PLoS One.* 2021;16(7):e0254638. doi: 10.1371/journal.pone.0254638
22. **Mavragani A, Gkillas K.** COVID-19 Predictability in the United States using Google Trends Time Series. *Sci Rep.* 2020;10(1), doi: 10.1038/s41598-020-77275-9

23. **Alshahrani R, Babour A.** An infodemiology and infoveillance study on COVID-19: analysis of Twitter and Google trends. *Sustainability.* 2021;13(15):8528. doi: 10.3390/su13158528

24. **Zhang X, Saleh H, Younis EMG,** *et al.* Predicting coronavirus pandemic in real-time using machine learning and big data streaming system. *Complexity.* 2020. doi: 10.1155/2020/6688912

25. **James G, Witten D, Hastie T,** *et al.* Tree-based methods. In: *An Introduction to Statistical Learning: with Applications in R.* 2nd ed. Springer; 2021.

26. **Cornelius E, Akman O, Hrozencik D.** COVID-19 mortality prediction using machine learning-integrated random forest algorithm under varying patient frailty. *Mathematics.* 2021;9(17):2043. doi: 10.3390/math9172043

27. **Iwendi C, Bashir AK, Peshkar A,** *et al.* COVID-19 patient health prediction using boosted random forest algorithm. *Front Public Health.* 2020;8:357. doi: 10.3389/fpubh.2020.00357

28. **Rozin, P, Royzman, EB.** Negativity bias, negativity dominance, and contagion. *Personal Soc Psychol Rev.* 2001;5(4):296-320. doi: 10.1207/S15327957PSPR0504_2

29. **Baumeister RF, Bratslavsky E, Finkenauer C,** *et al.* Bad is stronger than good. *Rev Gener Psychol.* 2001; 5(4):323-370. doi: 10.1037/1089-2680.5.4.323

30. **Baucum M, Cui J and John RS.** Temporal and geospatial gradients of fear and anger in social media responses to terrorism. *ACM Transactions on Social Computing.* 2020; 2(4):16, 1-16. doi.org/10.1145/3363565

31. **Yousefinaghani S, Dara R, Mubareka S,** *et al.* Prediction of COVID-19 waves using social media and Google search: a case study of the US and Canada. *Front Public Health.* 2021;9:656635. doi: 10.3389/fpubh.2021.656635

32. **Ridhwan KM, Hargreaves CA.** Leveraging Twitter data to understand public sentiment for the COVID-19 outbreak in Singapore. *Int J Inf Manag Data Insights.* 2021; 1(2):100021. doi: 10.1016/j.jjimei.2021.100021

33. **Israelashvili J.** More positive emotions during the COVID-19 pandemic are associated with better resilience, especially for those experiencing more negative emotions. *Front Psychol.* 2021;12:648112.

34. **Zepecki A, Guendelman S, DeNero J,** *et al.* Using application programming interfaces to access Google Data for health research: protocol for a methodological framework. *JMIR Res Protoc.* 2020;9(7):e16543. doi: 10.2196/16543

35. **Kim H, Jang SM, Kim S-H,** *et al.* Evaluating sampling methods for content analysis of Twitter data. *Soc Media Soc.* 2018. doi: 10.1177/2056305118772836

36. **Twitter.** Developer platform. Advanced filtering with geo data. Accessed June 27, 2023. https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data

37. **Kang H.** The prevention and handling of the missing data. *Korean J Anesthesiol.* 2013;64(5):402-406. doi: 10.4097/kjae.2013.64.5.402

38. **Breiman L.** Random forests. *Mach Learn.* 2001;45:5-32.

39. **Hastie T, Tibshirani R, Friedman J.** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. Springer; 2016

40. **Turk PJ, Tran TP, Rose GA,** *et al.* A predictive internet-based model for Covid-19 hospitalization census. *Sci Rep.* 2021;11(1):5106. doi: 10.1038/s41598-021-84091-2