healthcare

MDPI

Article

A Comparative Study on Distancing, Mask and Vaccine Adoption Rates from Global Twitter Trends

Satyaki Roy [1,*] and Preetam Ghosh [2]

[1] Department of Genetics, University of North Carolina, Chapel Hill, NC 27599, USA
[2] Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA; pghosh@vcu.edu
* Correspondence: satyakir@unc.edu

Abstract: COVID-19 is a global health emergency that has fundamentally altered human life. Public perception about COVID-19 greatly informs public policymaking and charts the course of present and future mitigation strategies. Existing approaches to gain insights into the evolving nature of public opinion has led to the application of natural language processing on public interaction data acquired from online surveys and social media. In this work, we apply supervised and unsupervised machine learning approaches on global Twitter data to learn the opinions about adoption of mitigation strategies such as social distancing, masks, and vaccination, as well as the effect of socioeconomic, demographic, political, and epidemiological features on perceptions. Our study reveals the uniform polarity in public sentiment on the basis of spatial proximity or COVID-19 infection rates. We show the reservation about the adoption of social distancing and vaccination across the world and also quantify the influence of airport traffic, homelessness, followed by old age and race on sentiment of netizens within the US.

Keywords: COVID-19; machine learning; tweets; sentiment analysis; adoption rates; socioeconomic

check for updates

Citation: Roy, S.; Ghosh, P. A Comparative Study on Distancing, Mask, and Vaccine Adoption Rates from Global Twitter Trends. *Healthcare* **2021**, *9*, 488. https://doi.org/10.3390/healthcare9050488

Academic Editor: Tingshao Zhu

Received: 10 March 2021
Accepted: 19 April 2021
Published: 21 April 2021

1. Introduction

The course and long-term impact of COVID-19 on human life will be determined largely by public attitude [1]. Since its inception in 2020, COVID-19 has left an indelible imprint on human existence and laid bare the socioeconomic, political, and attitudinal limitations of modern man [2]. COVID-19 has attracted the unprecedented attention of the international leadership because it has adversely affected both developed and developing nations alike, threatening to undo the years of sustainable growth especially in the least developed nations [3]. Epidemiologists agree that the adverse effects of future outbreaks can be contained by encouraging behavioral changes on the part of governments and individuals [4].

In the early stages of research on development, clinical trials and distribution of large-scale immunization measures, and social distancing continues to be a reliable strategy to curb infection spread and save lives [5]. The downside of the lockdown and international travel bans is the impending economic downturn and stock market decline [6]. This all-round existential threat prompted the scientific community constituting epidemiologists, sociologists, medical practitioners, and computer scientists to study the different aspects of the pandemic [7]. Several machine learning (ML) models are being created using the available clinical and epidemiological information to predict the course of the pandemic [8]. Specifically, ML models built around clinical data [8] are predicting the at-risk population [9,10], while epidemiological ML models are estimating the dynamics of contagion spread. Existing mechanistic, curve-fitting models [11], and supervised and unsupervised ML models [12] are bringing new insights into COVID spread dynamics. Khan et al. performed regression analysis, cluster analysis, and principal component analysis on Worldometer infection count data to study how variation in large-scale testing affects infection counts [13]. Roy et al. utilized regression studies to report pre-lockdown factors that impact post-lockdown contagion [14] and topic modeling to find the least and most affected economic sectors in the US [15]. They also proposed dynamic lockdown and mobility management strategies to curb contagion on the basis of economic and epidemiological profiles [16,17].

The sentiment about the COVID-19 mitigation will determine the strategies of effective policies and regulations to curb new waves or future outbreaks. Geldsetzer created a knowledge-base of perception of 80,000 volunteers from the US and UK, by performing an online survey [18]. Samuel et al. showed that Naive Bayes can achieve high sentiment prediction on COVID related tweets [19]. They outlined how policymakers and corporate houses will benefit from gauging the changing perceptions in the post-COVID world. They also explored the public sentiment about reopening in the new normal future [20]. Medford et al. applied topic modeling to show the sentiments about COVID-related topics, such as quarantine, transmission, and prevention, etc [21]. Paul presented an analysis on public sentiments over the opinions of social commentators and scientific communities about the origin of the virus [20]. It is imperative to suss the possible link between socioeconomic, geographical, political, and demographic factors of countries or US states and its public perception about adopting the emerging mitigation measures.

There have been efforts in leveraging network science to study the impact of public reservation over mask-wearing, vaccination, etc. Dinh et al. explored the relationship between the dynamics of virus and COVID information spread in social networks to show that they are interrelated but not identical [22]. Curiel compared the vaccination strategies against anti-vaccination views using diffusion models on different network topologies. They reported that reducing long distance trips or not allowing hubs (i.e., highly connected people in the social network) may dampen spread despite limited vaccine supply or anti-vaccination views [23]. Chung et al. explored the possibility that communication networks among health professionals can drive knowledge diffusion regarding COVID-19 mitigation [24]. Fagiolo studied the changing social network structure as more people are compelled to isolate due to the lockdown, linking disease diffusion and social network attributes [25]. Zhang et al. built a hidden multiplex network to quantify the interaction between information dissemination and virus spread [26].

Contributions

In this work, we carry out a comprehensive analytical study on the tweets about COVID-19 to identify the relationship between public sentiment and the socioeconomic, demographic, political and epidemiological features of a given geographical region (i.e., both country and US state). We apply supervised machine (and deep) as well as unsupervised learning approaches to learn the opinions about the adoption of social distancing, masks, and vaccination as a means to curb pandemic spread. Our study reveals the overall absence of variation in public sentiment on the basis of spatial proximity or COVID-19 infection rates. Furthermore, we observe a reluctance to recognize vaccination as a COVID-19 mitigation strategy. Specifically, a time-series analysis suggests a gradual decline in positive sentiments on social distancing and vaccination, while an increase in confidence on the use of masks. Our regression analysis elucidates that the US states with (1) high age and ethnicity (but not gender) affect public sentiment in the US, and (2) busy airports having a cosmopolitan population seem more accepting of the scientific research about vaccination and mask-wearing, and (3) a larger homeless population are stronger advocates of mask-wearing.

2. Materials and Methods

Here, we discuss the dataset utilized in the study, machine learning models for sentiment analysis and the performance metrics.

2.1. Dataset

2.1.1. COVID Tweet

We consider over 400,000 tweets between March to December 2020 from around the world. The tweets associated with COVID-19 were extracted in two steps. First, we referred to the repository that leveraged the Twitter's streaming API to amass tweet IDs (see https://github.com/echen102/COVID-19-TweetIDs, accessed on 20 February 2020 and [27]). Specifically, this repository contains text files with IDs (corresponding to COVID-19 related tweets) organized into folders in monthly chronological order. Second, we approximately collected the tweets by utilizing the first two text files in each folder (corresponding to the first day of each month) in conjunction with the Python TweePy API [28]. Tweets that were not posted in the English language (35.22% of total tweets) were deleted after posting were not included in our dataset. Each tweet has the following fields: {*ID*, *date*, *tweet*, *location*}. We considered tweets that relate to the following three COVID-19 keywords, namely *mask*, *vaccine*, and *distancing*. A tweet is about a keyword if it contains any of the words mentioned below.

- **Mask.** mask, masking, mask-wearing, wear, covering, face
- **Vaccine.** vaccine, vaccination, immune, immunity, dose, injection, inoculate, shot
- **Distancing.** lockdown, distancing, distance, mobility, isolation, social distance, contact, quarantine

Evidently, a tweet can be linked with multiple keywords and will therefore be counted several times. The world tweet dataset is available on GitHub (https://github.com/satunr/COVID-19/blob/master/Sentiment/final_world.csv.zip, accessed on 20 February 2020).

2.1.2. Latitude–Longitude Coordinates

The coordinates of the US states and countries are acquired from the Google Public Data Explorer [29]. It provides public data and forecasts (in a format called Dataset Publishing Language) in collaboration with several international organizations. The coordinates of the states and countries are shared in https://github.com/satunr/COVID-19/blob/master/Sentiment/Coordinate.xlsx, accessed on 20 February 2020.

2.1.3. Population Counts

The population of US states and countries is shared in GitHub (see https://github.com/satunr/COVID-19/blob/master/Sentiment/Population.xlsx, accessed on 20 February 2020).

2.1.4. Infection Counts

The infection counts for the US states are acquired from the data released by New York times (https://github.com/nytimes/covid-19-data, accessed on 20 February 2020). The infection count for the countries is taken from the COVID-19 Data Repository by the Center for Systems Science and Engineering at Johns Hopkins University (see https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv, accessed on 20 February 2020).

2.1.5. Red vs. Blue States

We create lists of likely red and blue states as per the projection of the outcome of the 2020 US presidential elections from NPR news [30] (see https://github.com/satunr/COVID-19/blob/master/Sentiment/red-blue.txt, accessed on 20 February 2020).

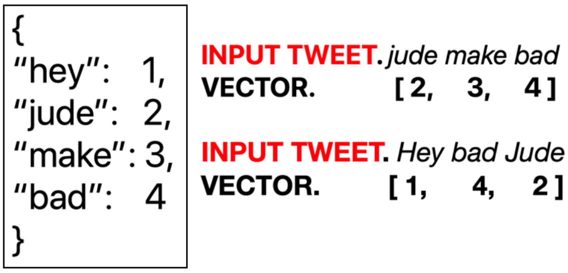
2.1.6. Demographic and Socioeconomic Features from the States of USA

We compiled this dataset from different open-access data portals. Each of the individual open-source datasets belonging to the 50 different US states and also the consolidated overall dataset are available on GitHub at the following link: https://github.com/satunr/COVID-19/tree/master/US-COVID-Dataset, accessed on 20 February 2020. We briefly explain the features compiled in this dataset below:

- *Age group* (total range: 0–80+ years) in groups of 4 years and also normalized them based on the population [31] (filename: source/Data_age.xlsx, feature name: age_to_, Norm_to_, e.g., age8to12 or Norm8to12); we have also grouped them in classes of 20 years later on for some of the analysis.
- *Percentage change in Gross Domestic Product* [32] It is measured in all 50 states and District of Columbia in 2020. (filename: source/GDP.xlsx, feature name: GDP).
- *Traffic or airport activity* quantifies the airport passenger traffic; it is also normalized by total airport traffic across all 50 US states [33] (filename: source/Data_airport.xlsx, feature name: Busy airport score, Normalized busy airport).
- *Gender* is the *male* and *female* fraction of the total population [34] (filename: source/Data_gender.csv, feature name: Male, Female).
- *Population density* is simply the ratio between the population of a state and its total area [31] (filename: source/Data_population.csv, feature name: Population, Area, Population Density).
- *Ethnicity* denotes the fraction(s) of total population that are *Asian*, *Hispanic*, *black*, and *white*; we did not consider the other smaller ethnicities [35] (filename: source/Data_ethnic.csv, feature name: Asian, Hispanic, Black and White).
- *Homeless* quantifies the total number of homeless individuals in each US state [36] (filename: source/Data_homeless.xlsx, feature name: Homeless). Similarly, for each state, the normalized homeless population designates the ratio between its homeless population and total population.
- *Healthcare index* is formally measured by the Agency for Healthcare Research and Quality (AHRQ). The quantification is done considering (1) the care setting (e.g., hospitals and nursing homes), (2) clinical area (e.g., caring for patients having diabetes, cancer), and (3) care type (e.g., chronic, preventive) [37] (filename: source/Data_health.xlsx, feature name: Health).

2.1.7. Preprocessing

We utilize the Python *Natural Language Toolkit* (NLTK) [38]—which is a standard library to process human language data—to acquire a labeled dataset of positive and negative tweets. In order to train a classifier, the labeled tweets are tokenized and stopwords are eliminated (Figure 1).



```
{
"hey":  1,
"jude": 2,
"make": 3,
"bad":  4
}
```

```
INPUT TWEET. jude make bad
VECTOR.      [ 2,  3,  4 ]

INPUT TWEET. Hey bad Jude
VECTOR.      [ 1,  4,  2 ]
```

Figure 1. Vectorization of tweet data: Each token is assigned an ID, and each tweet is converted into a vector of token IDs.

2.2. Sentiment Analysis Approaches

We discuss the machine learning approaches and performance metrics employed towards sentiment analysis.

2.2.1. Machine Learning Approaches

- *K-fold cross validation.* It is a standard approach to test the efficacy of a classification model [39]. The labeled training dataset is split into K partitions. The classifier is trained on all possible $K-1$ partitions and tested on the K-th partition. The overall performance of the estimator is a mean of the accuracy of the predictions from the K folds. We used the Python *Scikit-learn library* [40] to perform cross-validation.

- *Naive Bayes (NB).* It is a class of fast, probabilistic learning techniques that apply the Bayes' theorem to assign labels to the data points [41]. We use NB to learn the sentiments in an iterative approach. We train the NB classifier with the labeled NLTK tweet dataset and iteratively predict the label of p% unlabeled COVID tweet dataset. At each step, we apply K-fold cross validation (see Section 2.2.1) to monitor the accuracy of the integrated dataset.
- *Convolutional Neural Networks (CNN).* It is a class of deep learning which can assign varying importance to different aspects of the input data. This helps them distinguish these aspects from one another [42]. After the processing step (see Section 2.1.7) is complete, we assign a unique ID to the meaningful tokens and *vectorize* each labeled tweet as follows. Each labeled tweet is filtered to preserve L most common words and converted to a vector of token IDs (see Section 1), allowing us to preserve the relative order of the tokens. The CNN classifier is then trained on the (tweet vector, sentiment) pair and the accuracy, estimated by cross-validation, is recorded. At the end of the training stage, a tweet is associated with a label 2 or 0, representing positive or negative sentiments, respectively.
- *Multiple regression (MR).* Considering a dependent and independent variable B and A respectively, their linear association designated by the function $B = g(A)$ is captured by the statistical measure of MR. The linear association is generated by MR in the following form: $\hat{B} = \beta(0) + \beta(f_1)A(f_1) + \beta(f_2)A(f_2) + \cdots + \epsilon$; here, ϵ is the error term, $\beta(0)$ the intercept, and the coefficient $\beta(f_i)$ quantifies how feature f_i contributes to b, which is the dependent variable.
- *Curve fitting.* It is an approach that estimates the coefficients of a polynomial of degree d that minimizes the least squares error for input vectors A and B [43]. We utilize the Python NumPy library [44] to achieve curve-fitting.

CNN has been used to analyze image data, but also learn the relationship of text features in the absence of the knowledge of the entire context [45]. Like CNN, Naive Bayes has also emerged as a reliable approach for sentiment analysis. However, unlike CNN, Naive Bayes offers quick training with fewer parameters, yet it yields fairly accurate performance [46]. We use the curve-fitting to learn the evolving trends in the COVID-19 adoption rates, while the regression studies is intended to quantify the contributions of the several socioeconomic and demographic factors towards the adoption rates.

2.3. Performance Metrics

We also utilize the following metrics in our study.

2.3.1. Accuracy

This metric is expressed as the ratio between the observations that were correctly predicted to the total number of observations [47]. In other words, in multilabel classification problems, it quantifies the fraction of agreements between the predicted labels and the actual ones:

$$Acc = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \tag{1}$$

Here, TN, TP, FN, and FP respectively designate the true negatives, true positives, false negatives, and false positives.

2.3.2. Euclidean Distance

It is the length of a line segment between any two given points p and q in a space of m dimensions. It is calculated as:

$$D = \sqrt{\sum_{i}^{m} (p_i - q_i)^2} \tag{2}$$

2.3.3. One Sample *t*-Test

This is a statistical testing that compares the mean of a sample data to a known value. T score captures the number of standard deviations by which the sample mean deviates from known value [48].

2.3.4. Pearson Correlation Coefficient

This statistical metric captures the linear association between two different variables where -1 and 1 designate strong negative and positive correlation, respectively, and 0 designates that the variables that are not correlated. If A and B are the two variables, the Pearson correlation coefficient is given by $\frac{\text{cov(A,B)}}{\sigma_A \times \sigma_B}$, where $\text{cov}(\text{A}, \text{B})$ and σ_A denote the covariance matrix and standard deviations, respectively [49].

Accuracy has been employed to quantify the ability of the machine learning classifier to recognize the public sentiment. The Euclidean distance, in conjunction with a one-sample *t*-test and Pearson correlation coefficient, helps study the correlation between spatial proximity between places (states or nations) and similarity in public perceptions.

3. Results

This section comprises the following subsections: selection of classifier to learn tweet sentiments, effect of spatial proximity, infection rates and political ideology on sentiment, evolution of opinions about COVID mitigation measures, and the effect of socioeconomic, demographic, and epidemiological factors on sentiment.

3.1. Selection of Classifier

We apply $K-$fold cross-validation on the Naive Bayes and CNN classifiers and determine their respective accuracy in predicting tweet sentiment. Figure 2 shows the varying cross-validation accuracy using CNN on the labeled NLTK tweet dataset (see Section 2.2 for details).

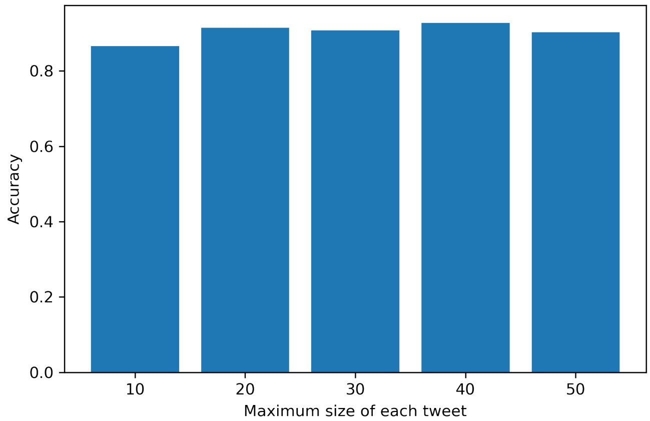


Figure 2. Accuracy of $K-$fold cross-validation on labeled NLTK tweet data using CNN with varying length of tweet vector.

We vary the number of tokens in each tweet (also measured as the length of each tweet vector) $L = 10, 20, \cdots, 50$. However, the overall accuracy is approximately 85%.

A similar $K-$fold cross-validation experiment using the Naive Bayes (NB) classifier on the labeled NLTK tweet data yields an approximate sentiment prediction accuracy of 90%. As described in Section 2.2.1, we subsequently train NB with the augmented dataset (comprising labeled NLTK tweet dataset and part of COVID tweet data) and iteratively predict the label of $p = 5\%$ yet *unlabeled* COVID tweets. Applying $K-$fold cross validation at each step, NB achieves an overall accuracy of around 95% on the combined dataset (as shown in Figure 3).

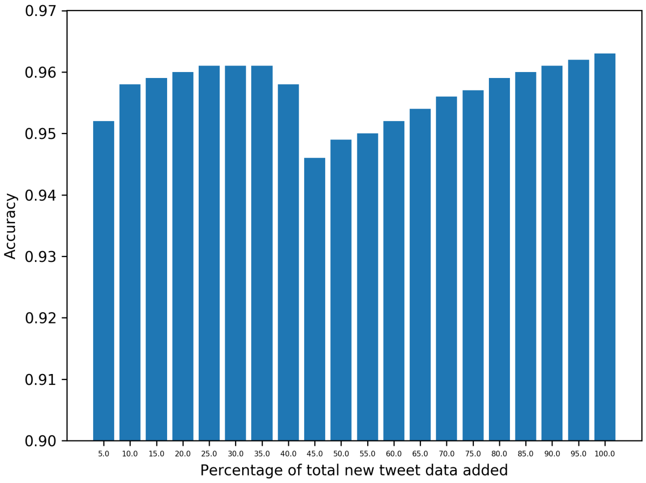


Figure 3. Accuracy of $K-$fold cross-validation (using Naive Bayes classifier) on augmented tweet data constituting labeled NLTK tweets and $p = 5\%, 10\%, \ldots,$ COVID-19 tweets.

3.2. *Correlation between Geographical Proximity and Sentiment*

We utilize the Python library for geocoding, called GeoPy [50], to identify the latitude-longitude of the countries and US states. The details of this step are covered in Section 2.1.2. For each location (i.e., US state, country) c, we gauge the fraction of positive tweets in the region F ($F > 0.55$). For any pair of regions c_i and c_j, we calculate the Pearson's coefficient (see Section 2.3.4) between the geographical proximity (measured in terms of Euclidean distance defined in Section 2.3.2) and the absolute difference between their fraction of positive tweets $F(c_i)$ and $F(c_j)$, i.e., $|F(c_i) - F(c_j)|$.

In Figure 4a,b, we place the countries and US states on their latitude–longitude, where the F is shown in different colors and annotated with the (t-statistic, p-value) for the hypothesis that the mean positivity in tweet sentiment is >0.5. The positive tweets and total tweets of the places with the least p-value and $F > 0.55$ are summarized in Table 1. Note that countries like Azerbaijan, Bouvet Island, Gabon, Kyrgyzstan, Liechtenstein, Marshall Islands, Mauritius, Micronesia, Pitcairn Islands, Saint Vincent and the Grenadines, Solomon Islands, Togo, Tuvalu, and Vatican City that have F$\sim$1.0 are not statistically significant as few tweets originate from these places. Pearson's correlation, p-value, for world and US are $(0.03, 2.3010^{-08})$ and $(-0.01, 0.60)$, respectively, suggesting the overall lack of correlation between geographical proximity and public sentiment among the countries and US states.

Table 1. Positive tweets, total tweets of countries and states with the least p-value and fraction of positive tweets $F > 0.55$.

Place	Positive	Total Tweets
Argentina	359	520
Brazil	962	1651
India	4796	8051
Indonesia	794	1317
Peru	216	307
Guam	39	64
Vermont	169	293
North Dakota	166	289
Rhode Island	302	542
West Virginia	175	318

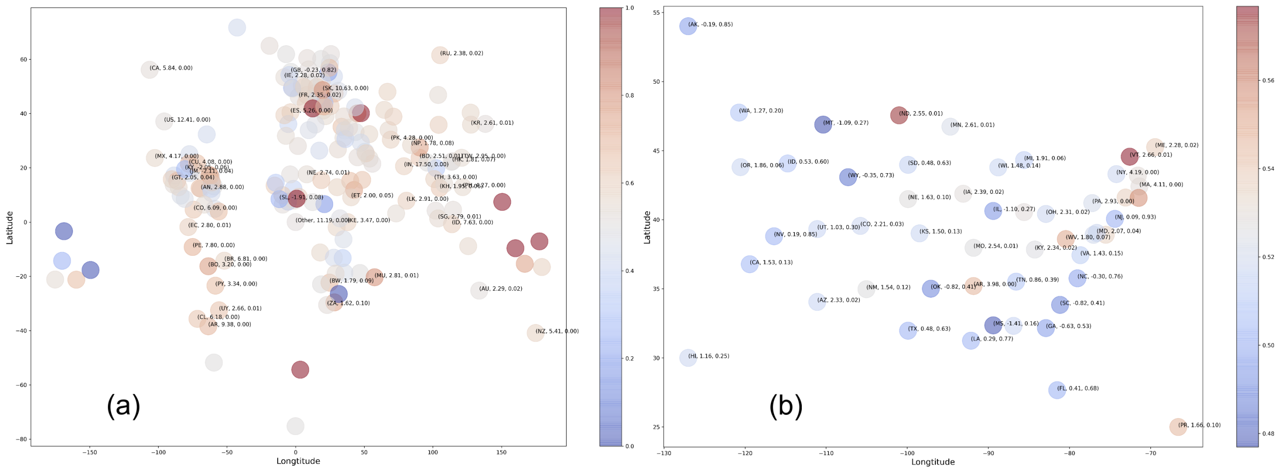


Figure 4. Visualization of the (**a**) countries and (**b**) US states based on latitude–longitude. The colors represent the fraction of positive tweets and annotations show (t-statistic, p-value) for hypothesis that the mean positivity in tweet sentiment is >0.5.

3.3. *Effect of Sentiments on COVID-19 Infection Rates*

We rank the countries and US states c in the increasing order of the positive sentiments $F(c)$ corresponding to the tweet set on vaccine, mask, and social distancing. For each region (i.e., country and state) and tweet set, we calculate the total COVID-19 infection numbers spanning between the earliest tweet date of the tweet set and normalize it by the population of that place (see Section 2.1.3 for details). Figure 5a–f show the plots corresponding to vaccine, mask, and social distancing related tweets and infection numbers for the countries and states, respectively. The results show that USA, India, and Brazil (colored red) show a high normalized infection count, but they are spread all across the spectrum of countries ranked by sentiment. Similarly, Alaska, Delaware, and Maine (colored red) are amongst the states with the highest scaled infection count between the corresponding tweet dates. Interestingly, these states exhibit the least positive sentiment in case of vaccine (Figure 5d) and mask-wearing (Figure 5e), suggesting that the reluctance of the public in adopting COVID-19 mitigation measures may be a contributor to the overall contagion within the US. However, no such correlations can be established from the social distancing plot (Figure 5f).

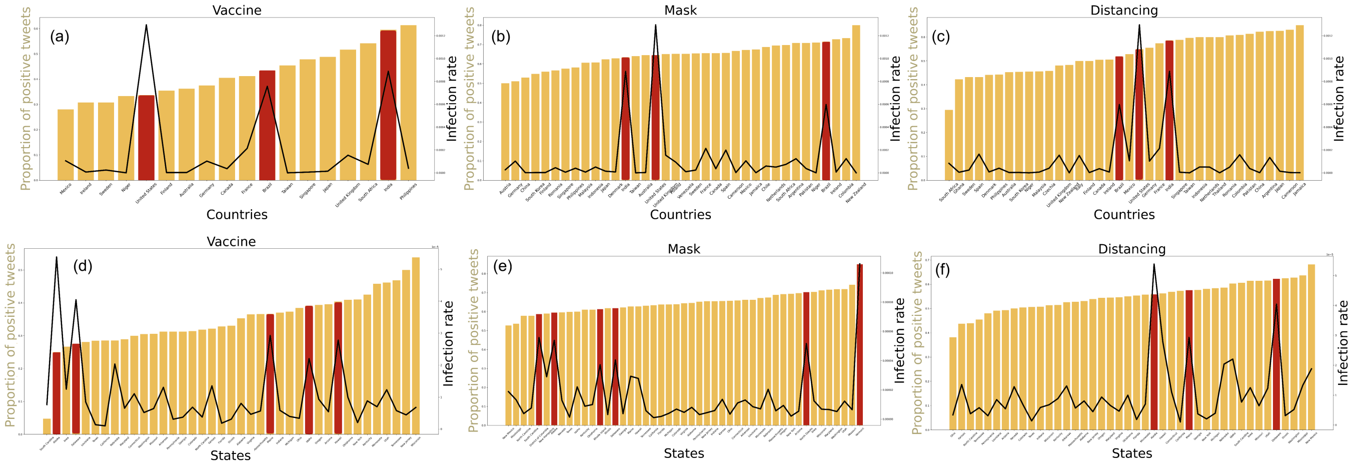


Figure 5. Fraction of positive tweets on vaccine, mask and social distancing from the (**a**–**c**) countries and (**d**–**f**) US states and their total infection numbers (between the tweet period) normalized by their population count.

3.4. Evolution of Public Sentiment over COVID-19

We isolate the tweets regarding (1) vaccine, (2) mask, and (3) social distancing related terms (refer Section 2.1.1 for the related terms of the three keywords). Since each tweet is associated with a date, we visualize the public sentiment (measured, once again, in terms of the fraction of positive tweets F) against the time-wise evolution of the three COVID-19 keywords.

Figures 6 and 7 show that there exist similar trends in public tweets from the world and US about vaccine, mask, and social distancing. In both cases, we see an increasing positivity over mask-wearing with F∼0.7 as of November 2020. We fit the curve to a polynomial of order 2 to show that the positivity in tweet sentiment over masking is expected to grow in 2021. Conversely, the trends are opposite in the case of vaccine and social distancing related tweets, as the adoption rates of these keywords continue to decline over time. It is evident that the netizens, mostly comprising a younger population, have either made light of the severity of the pandemic or been hesitant (a) to adopt social distancing and (b) about the efficacy of vaccines in combating pandemic spread.

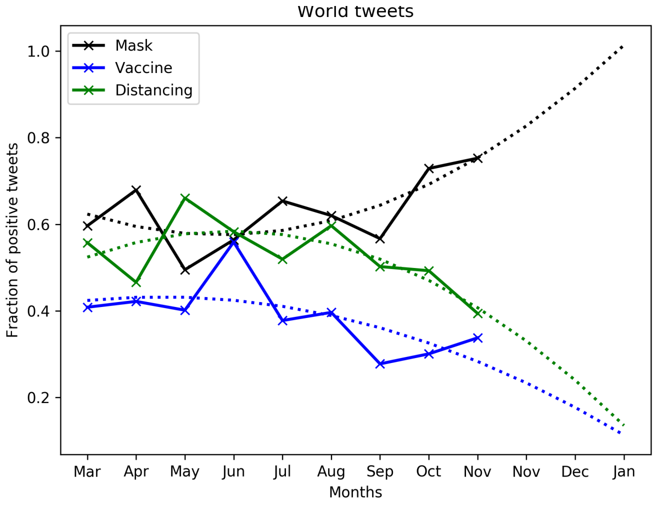


Figure 6. Evolution of public sentiment in world tweets about COVID-19 keywords, vaccine, mask, and distancing over time.

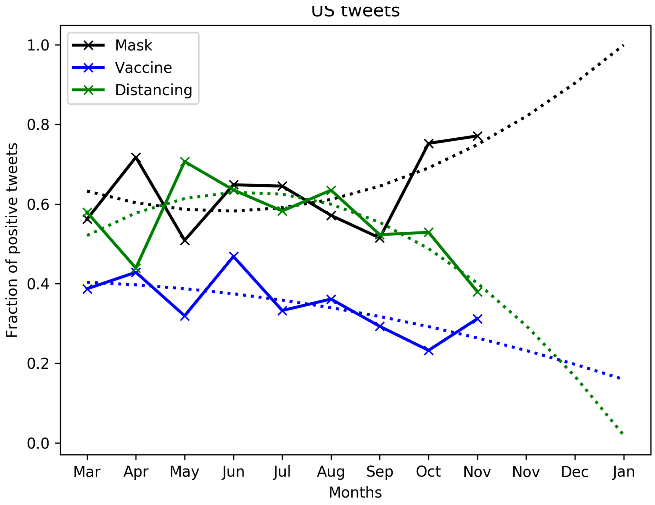


Figure 7. Evolution of public sentiment in US tweets about COVID-19 keywords, vaccine, mask, and distancing over time.

3.5. Effect of Political Ideology on Sentiments

We plot the fraction of positive tweets $F(c)$ originating from each state c and estimate the mean and standard deviation in $F(c)$ for the red and blue states (reported in Section 2.1.5). Figure 8 shows that $F(c)$ for the red and blue states are comparable. The blue states tend to exhibit a marginally higher (and notably lesser variation in) positivity in COVID-19 related tweets.

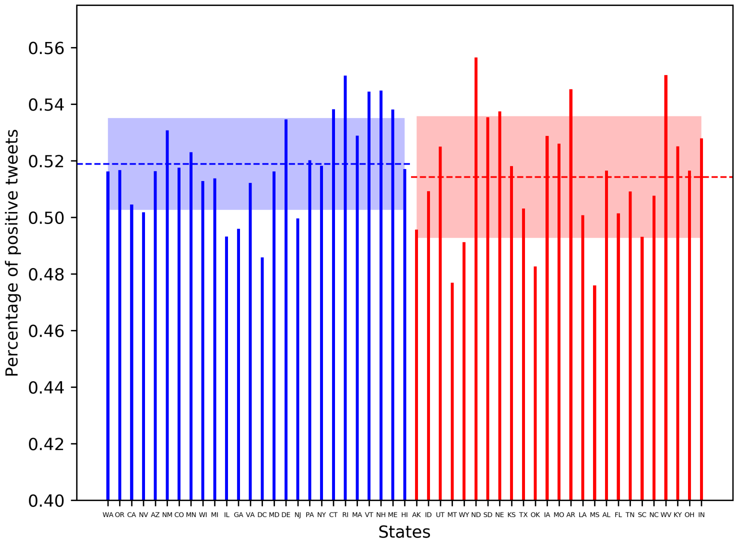


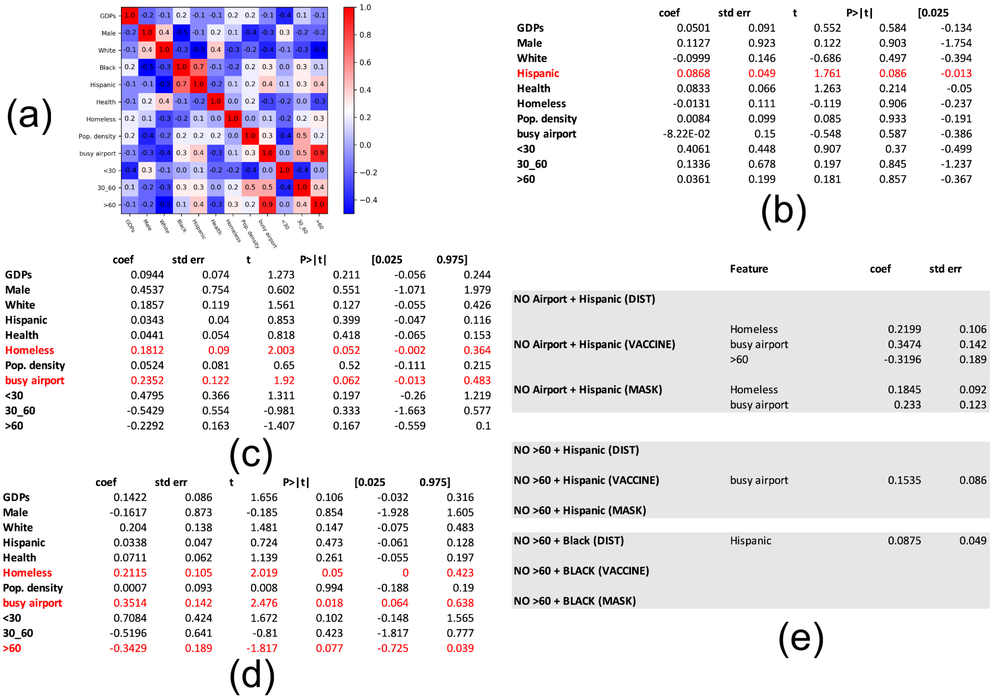
Figure 8. Mean and standard deviation in the fraction of positive sentiment tweets F for the blue and red states in USA.

3.6. Effect of Socioeconomic and Demographic Features on Sentiments on Adoption of COVID Mitigation Measures

We use multiple regression (see Section 2.2.1) to gauge the possible effect of the socioeconomic and demographic features of US states on the adoption rates of COVID-19 mitigation measures such as *vaccine*, *mask*, and *social distancing*. We consider a feature set comprising GDP, gender, ethnicity, health, normalized homeless, and three age brackets (0–30, 31–60 and >60) from the dataset (see Section 2.1.6 for details on each factor). For the labeled US tweet dataset, we consider the subsets of tweets containing (1) all terms, (2) vaccine, (3) mask, and (4) distancing related terms (refer to Section 2.1.1 for the related terms of the three keywords). Each feature value is normalized by the maximum value of that feature across all US states. For each state c, the dependent variable in the regression study is the fraction of positive tweets $F(c)$ among the selected tweet subset.

Beginning with the feature set GDP, male, white, black, hispanic, health index, homeless, population density, airport traffic, and age groups <30, 30–60, >60, we measure the Pearson correlation between every pair of features. The features with a correlation equal to or in excess of 0.7 are not considered in the same regression analysis. Figure 9a shows two such pairs of correlated features, namely (age group $>$ 60, airport traffic) and (black, hispanic). Thus, in the regression analysis, we consider the four configurations without the following feature pairs: (1) (airport, black), (2) (airport, Hispanic), (3) (age $>$ 60, hispanic) and (4) (age $>$ 60, black), ensuring that the remaining features in each configuration are not highly correlated. We carry out the regression analysis for all four configurations and record the significant features (p-value $<$ 0.1).

We show the complete regression tables for social distancing, mask-wearing, and vaccination for configuration (1) in Figure 9b–d, respectively. Note that homeless and airport traffic emerge as the most significant features contributing positively to F. Hispanic and age $>$ 60 contribute positively and negatively to public perception around distancing and vaccination, respectively. Figure 9e shows the statistically significant features for configurations (2)–(4), where once again homeless and airport traffic emerge as the highest positive contributors. This suggests that the states with a high proportion of homeless are respectful of the effects of the pandemic and particularly how mask-wearing can mitigate its ill-effects. In addition, people living in states with high airport traffic, representing the cosmopolitan section of the US population, are clued in to the scientific advancements to curb spread. Similarly, the places with a relatively younger population (i.e., age $\leq$ 60) are flouting the COVID-19 related restrictions.



	GDPs	Male	White	Black	Hispanic	Health	Homeless	Pop. density	busy airport	<30	30_60	>60
GDPs	1.0	-0.2	-0.1	0.2	-0.1	-0.1	0.2	0.2	-0.1	-0.4	0.1	-0.1
Male	-0.2	1.0	0.4	-0.5	-0.1	0.2	0.1	-0.4	-0.3	0.3	-0.2	-0.2
White	-0.1	0.4	1.0	-0.3	-0.5	0.4	-0.3	-0.2	-0.4	-0.1	-0.3	-0.5
Black	0.2	-0.5	-0.3	1.0	0.7	-0.1	-0.2	0.2	0.3	0.0	0.3	0.1
Hispanic	-0.1	-0.1	-0.5	0.7	1.0	-0.2	0.1	0.2	0.4	0.1	0.3	0.4
Health	-0.1	0.2	0.4	-0.1	-0.2	1.0	0.0	0.2	-0.3	-0.2	0.0	-0.3
Homeless	0.2	0.1	-0.3	-0.2	0.1	0.0	1.0	0.0	0.1	-0.2	0.2	0.3
Pop. density	0.2	-0.4	-0.2	0.2	0.2	0.2	0.0	1.0	0.3	-0.4	0.5	0.2
busy airport	-0.1	-0.3	-0.4	0.3	0.4	-0.3	0.1	0.3	1.0	0.0	0.5	0.9
<30	-0.4	0.3	-0.1	0.0	0.1	-0.2	-0.2	-0.4	0.0	1.0	-0.4	0.0
30_60	0.1	-0.2	-0.3	0.3	0.3	0.0	0.2	0.5	0.5	-0.4	1.0	0.4
>60	-0.1	-0.2	-0.5	0.1	0.4	-0.3	0.3	0.2	0.9	0.0	0.4	1.0

	coef	std err	t	P>|t|	[0.025
GDPs	0.0501	0.091	0.552	0.584	-0.134
Male	0.1127	0.923	0.122	0.903	-1.754
White	-0.0999	0.146	-0.686	0.497	-0.394
Hispanic	0.0868	0.049	1.761	0.086	-0.013
Health	0.0833	0.066	1.263	0.214	-0.05
Homeless	-0.0131	0.111	-0.119	0.906	-0.237
Pop. density	0.0084	0.099	0.085	0.933	-0.191
busy airport	-8.22E-02	0.15	-0.548	0.587	-0.386
<30	0.4061	0.448	0.907	0.37	-0.499
30_60	0.1336	0.678	0.197	0.845	-1.237
>60	0.0361	0.199	0.181	0.857	-0.367

	coef	std err	t	P>|t|	[0.025	0.975]
GDPs	0.0944	0.074	1.273	0.211	-0.056	0.244
Male	0.4537	0.754	0.602	0.551	-1.071	1.979
White	0.1857	0.119	1.561	0.127	-0.055	0.426
Hispanic	0.0343	0.04	0.853	0.399	-0.047	0.116
Health	0.0441	0.054	0.818	0.418	-0.065	0.153
Homeless	0.1812	0.09	2.003	0.052	-0.002	0.364
Pop. density	0.0524	0.081	0.65	0.52	-0.111	0.215
busy airport	0.2352	0.122	1.92	0.062	-0.013	0.483
<30	0.4795	0.366	1.311	0.197	-0.26	1.219
30_60	-0.5429	0.554	-0.981	0.333	-1.663	0.577
>60	-0.2292	0.163	-1.407	0.167	-0.559	0.1

	coef	std err	t	P>|t|	[0.025	0.975]
GDPs	0.1422	0.086	1.656	0.106	-0.032	0.316
Male	-0.1617	0.873	-0.185	0.854	-1.928	1.605
White	0.204	0.138	1.481	0.147	-0.075	0.483
Hispanic	0.0338	0.047	0.724	0.473	-0.061	0.128
Health	0.0711	0.062	1.139	0.261	-0.055	0.197
Homeless	0.2115	0.105	2.019	0.05	0	0.423
Pop. density	0.0007	0.093	0.008	0.994	-0.188	0.19
busy airport	0.3514	0.142	2.476	0.018	0.064	0.638
<30	0.7084	0.424	1.672	0.102	-0.148	1.565
30_60	-0.5196	0.641	-0.81	0.423	-1.817	0.777
>60	-0.3429	0.189	-1.817	0.077	-0.725	0.039

	Feature	coef	std err
NO Airport + Hispanic (DIST)			
NO Airport + Hispanic (VACCINE)	Homeless busy airport >60	0.2199 0.3474 -0.3196	0.106 0.142 0.189
NO Airport + Hispanic (MASK)	Homeless busy airport	0.1845 0.233	0.092 0.123
NO >60 + Hispanic (DIST)			
NO >60 + Hispanic (VACCINE)	busy airport	0.1535	0.086
NO >60 + Hispanic (MASK)			
NO >60 + Black (DIST)	Hispanic	0.0875	0.049
NO >60 + BLACK (VACCINE)			
NO >60 + BLACK (MASK)			

Figure 9. Regression analysis. (**a**) pairwise feature correlation; (**b**) complete regression table for feature set without airport and black features; significant features for regression of feature set without (**c**) airport and Hispanic) (**d**) age > 60 and hispanic, and (**e**) age > 60, black.

4. Discussion

While this study provides some valuable insights into the general attitude surrounding COVID-19 mitigation measures, it suffers from the usual challenges associated with any natural language processing based study. First, with regard to the sentiment analysis of the countries, the entire analysis is done for English tweets, which only accounts for 50% of global tweet dataset [51] in general and 64.78% of tweets in this study; therefore, it does not adequately represent the non-English speaking nations. However, we report the *p*-values in Section 3.2 to give the readers a sense of how reliable the sentiment for each country may be. Nonetheless, we find that the broad trends in sentiments are quite consistent across nations. Next, we show that the frequency of tweets across UK, USA vary greatly from countries like Togo and Solomon Islands, unavoidably skewing our measure of public perception. In addition, in many ways, it is more useful to know the sentiment of people hailing from smaller, developing nations. By the same token, the sentiments expressed by the netizens is not always the best representation of the worst-affected section of the society, such as the homeless or economically challenged. Finally, there is not enough data about the socioeconomic background of the individual issuing the tweet or his followers. This knowledge may present interesting understanding on the biases and the influence they may wield on public sentiment.

5. Conclusions

In this paper, we analyzed over 400,000 tweets from across the world to infer the evolution of public sentiments regarding the adoption of COVID-19 mitigation measures, namely vaccination, mask-wearing, and social distancing. We employ supervised machine learning methods in conjunction with natural language processing to identify the socioeconomic, demographic, epidemiological as well as political factors dictating public perception of the mitigation strategies.

Our analysis leads to several interesting findings. First, while there is a near-perfect balance between positive and negative sentiments globally, we see a lack of correlation between spatial proximity and public perception. Second, the blue states (in the USA) seem marginally more receptive to the COVID-19 mitigation measures, and the public at large seems to be responding positively to mask-wearing as a way of curbing spread. Third, our regression analysis reveals that the US states with a high number of homeless individuals show greater awareness of the perils of the pandemic and that high age and ethnicity (but not gender) are key predictors contributing to public attitudes about COVID-19.

Author Contributions: Conceptualization, S.R. and P.G.; methodology and software, S.R.; validation, P.G.; resources, P.G.; data curation, S.R.; writing—original draft preparation, S.R.; writing—review and editing, S.R. and P.G.; visualization, S.R.; supervision and project administration, P.G.; funding acquisition, P.G. Both authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by NSF grant CBET-1802588. The opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and the codes are available online at https://github.com/satunr/COVID-19, accessed on 20 February 2020.

Acknowledgments: The authors thank Ronojoy Dutta from Deep Run High School, Virginia, USA for his extensive help with data curation, software building, and visualization for this article. He helped with all aspects of the implementation of the proposed informatics pipeline as well as with interpretation of the results.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Sibley, C.G.; Greaves, L.M.; Satherley, N.; Wilson, M.S.; Overall, N.C.; Lee CH, J.; Milojev, P.; Bulbulia, J.; Osborne, D.; Milfont, T.L.; et al. Effects of the COVID-19 pandemic and nationwide lockdown on trust, attitudes toward government, and well-being. *Am. Psychol.* **2020**, *75*, 618–630. [CrossRef]
2. Coronavirus: What Have Been the Worst Pandemics and Epidemics in History? 2020. Available online: https://en.as.com/en/2020/04/18/other_sports/1587167182_422066.html (accessed on 20 February 2021).
3. COVID-19 and the Least Developed Countries. 2020. Available online: https://www.un.org/development/desa/dpad/publication/un-desa-policy-brief-66-covid-19-and-the-least-developed-countries/ (accessed on 20 February 2021).
4. Scudellari, M. How the pandemic might play out in 2021 and beyond. *Nature* **2020**, *584*, 22–25. [CrossRef] [PubMed]
5. Coronavirus, Social and Physical Distancing and Self-Quarantine. 2020. Available online: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/coronavirus-social-distancing-and-self-quarantine (accessed on 20 February 2021).
6. Ozili, P.; Arun, T. Spillover of COVID-19: Impact on the Global Economy. *SSRN Electron. J.* **2020**. [CrossRef]
7. Adhikari, S.; Meng, S.; Wu, Y.; Mao, Y.; Ye, R.; Wang, Q.; Sun, C.; Sylvia, S.; Rozelle, S.; Raat, H.; et al. Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. *Infect. Dis. Poverty* **2020**, *9*, 29. [CrossRef] [PubMed]
8. Wynants, L.; Calster, B.V.; Bonten, M.; Collins, G.S.; Smeden, M.V. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **2020**, *369*. [CrossRef] [PubMed]
9. Alimadadi, A.; Aryal, S.; Manandhar, I.; Munroe, P.; Joe, B.; Cheng, X. Artificial intelligence and machine learning to fight COVID-19. *Physiol. Genom.* **2020**, *52*, 200–202. [CrossRef] [PubMed]

10. Randhawa, G.; Soltysiak, M.; El Roz, H.; de Souza, C.; Hill, K.; Kari, L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLoS ONE* **2020**, *15*, e0232391. [CrossRef]
11. Holmdahl, I.; Buckee, C. Wrong but useful—What covid-19 epidemiologic models can and cannot tell us. *N. Engl. J. Med.* **2020**, *383*, 303–305. [CrossRef]
12. Wang, P.; Zheng, X.; Li, J.; Zhu, B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos Solitons Fractals* **2020**, *139*, 110058. [CrossRef] [PubMed]
13. Khan, N.; Naushad, M.; Fahad, S.; Faisal, S.; Muhammad, A. COVID-2019 and World Economy. *J. Health Econ. Forthcom.* **2020**. [CrossRef]
14. Roy, S.; Ghosh, P. Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking. *PLoS ONE* **2020**, *15*, e0241165. [CrossRef]
15. Roy, S.; Dutta, R.; Ghosh, P. Recreational and philanthropic sectors are the worst-hit US industries in the COVID-19 aftermath. *Soc. Sci. Humanit. Open* **2021**, *3*, 100098. [CrossRef]
16. Roy, S.; Dutta, R.; Ghosh, P. Towards Dynamic lockdown strategies controlling pandemic spread under healthcare resource budget. *Appl. Netw. Sci.* **2021**, *6*, 1–15. [CrossRef] [PubMed]
17. Roy, S.; Cherevko, A.; Chakraborty, S.; Ghosh, N.; Ghosh, P. Leveraging Network Science for Social Distancing to Curb Pandemic Spread. *IEEE Access* **2021**, *9*, 26196–26207. [CrossRef]
18. Geldsetzer, P. Knowledge and perceptions of COVID-19 among the general public in the United States and the United Kingdom: A cross-sectional online survey. *Ann. Intern. Med.* **2020**, *173*, 157–160. [CrossRef]
19. Samuel, J.; Ali, G.; Rahman, M.; Esawi, E.; Samuel, Y. Covid-19 public sentiment insights and machine learning for tweets classification. *Information* **2020**, *11*, 314. [CrossRef]
20. Samuel, J.; Rahman, M.; Ali, G.; Samuel, Y.; Pelaez, A.; Chong, P.H.J.; Yakubov, M. Feeling Positive About Reopening? New Normal Scenarios From COVID-19 US Reopen Sentiment Analytics. *IEEE Access* **2020**, *8*, 142173–142190. [CrossRef]
21. Medford, R.; Saleh, S.; Sumarsono, A.; Perl, T.; Lehmann, C. An "infodemic": Leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak. In *Open Forum Infectious Diseases*; Oxford University Press US: New York, NY, USA, 2020; Volume 7, p. ofaa258.
22. Dinh, L.; Parulian, N. COVID-19 pandemic and information diffusion analysis on Twitter. *Proc. Assoc. Inf. Sci. Technol.* **2020**, *57*, e252. [CrossRef]
23. Curiel, R.P.; Ramírez, H.G. Vaccination strategies against COVID-19 and the diffusion of anti-vaccination views. *Sci. Rep.* **2021**, *11*, 1–13.
24. Chung, M.; Dekker, D.; Gridley-Smith, C.; Dearing, J. Peer Reviewed: An Emergent Network for the Diffusion of Innovations Among Local Health Departments at the Onset of the COVID-19 Pandemic. *Prev. Chronic Dis.* **2021**, *18*, E19. [CrossRef]
25. Fagiolo, G. Assessing the Impact of Social Network Structure on the Diffusion of Coronavirus Disease (COVID-19): A Generalized Spatial SEIRD Model. *arXiv* **2020**, arXiv:2010.11212.
26. Zhang, X.; Zhang, Z.; Wang, W.; Hou, D.; Xu, J.; Ye, X.; Li, S. Multiplex network reconstruction for the coupled spatial diffusion of infodemic and pandemic of COVID-19. *Int. J. Digit. Earth* **2021**, *14*, 401–423. [CrossRef]
27. Chen, E.; Lerman, K.; Ferrara, E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health Surveill.* **2020**, *6*, e19273. [CrossRef] [PubMed]
28. Roesslein, J. Tweepy Documentation. 2009, 5. Available online: http://tweepy.readthedocs.io/en/v3 (accessed on 20 February 2021).
29. Google. Google Public Data Explorer. 2010. Available online: http://www.google.com/publicdata/home (accessed on 20 February 2021).
30. NPR Elections: Presidential Results. 2020. Available online: https://apps.npr.org/elections20-interactive/ (accessed on 20 February 2021).
31. Census, U.S. State Population by Characteristics: 2010–2019. 2019. Available online: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-detail.html (accessed on 20 February 2021).
32. Bureau of Economic Analysis. Gross Domestic Product by State. 2020. Available online: https://www.bea.gov/news/2021/gross-domestic-product-state-4th-quarter-2020-and-annual-2020-preliminary (accessed on 20 February 2021).
33. Wikipedia. List of the Busiest Airports in the United States. 2019. Available online: https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States (accessed on 20 February 2021).
34. KFF. Population Distribution by Gender. 2017. Available online: https://www.kff.org/other/state-indicator/distribution-by-gender/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D (accessed on 20 February 2021).
35. KFF. Population Distribution by Race/Ethnicity. 2018. Available online: https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?dataView=0&currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D (accessed on 20 February 2021).
36. Exchange, H. 2013 AHAR: Part 1—PIT Estimates of Homelessness in the U.S. 2013. Available online: https://www.hudexchange.info/resource/3300/2013-ahar-part-1-pit-estimates-of-homelessness/ (accessed on 20 February 2021).
37. Agency for Healthcare Research and Quality. Health Care Quality: How Does Your State Compare? Available online: https://www.ahrq.gov/data/infographics/state-compare-text.html (accessed on 20 February 2021).

38. Bird, S. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*; Association for Computational Linguistics (ACL): Sydney, Australia, 2006; pp. 69–72.
39. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
40. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
41. Rish, I. An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop Empir. Methods Artif. Intell.* **2001**, *3*, 41–46.
42. Albawi, S.; Mohammed, T.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6.
43. Lesh, F. Multi-dimensional least-squares polynomial curve fitting. *Commun. ACM* **1959**, *2*, 29–30. [CrossRef]
44. McKinney, W. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*; O'Reilly Media, Inc.: Newton, MA, USA, 2012.
45. Liao, S.; Wang, J.; Yu, R.; Sato, K.; Cheng, Z. CNN for situations understanding based on sentiment analysis of twitter data. *Procedia Comput. Sci.* **2017**, *111*, 376–381. [CrossRef]
46. Sunarya, P.; Refianti, R.; Mutiara, A.; Octaviani, W. Comparison of Accuracy between Convolutional Neural Networks and Naïve Bayes Classifiers in Sentiment Analysis on Twitter. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 77–86. [CrossRef]
47. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning, Haifa, Israel, 21–25 June 2010; pp. 233–240.
48. Ross, A.; Willson, V. One-sample *t*-test. In *Basic and Advanced Statistical Tests*; Brill Sense: Leiden, The Netherlands, 2017; pp. 9–12.
49. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing;* Springer Topics in Signal Processing; Springer: Berlin/Heidelberg, Germany, 2009; Volume 2. [CrossRef]
50. GeoPy: Geocoding Library for Python. 2020. Available online: https://github.com/geopy/geopy (accessed on 20 February 2021).
51. Alowibdi, J.; Buy, U.; Yu, P. Empirical evaluation of profile characteristics for gender classification on twitter. In Proceedings of the 2013 12th International Conference on Machine Learning and Applications, Miami, FL, USA, 4–7 December 2013; Volume 1, pp. 365–369.