Forecasting COVID-19 Vaccination Rates using Social Media Data

Xintian Li
xli71@tulane.edu
Tulane University
New Orleans, Louisiana, USA

Aron Culotta
aculotta@tulane.edu
Tulane University
New Orleans, Louisiana, USA

ABSTRACT

The COVID-19 pandemic has had a profound impact on the global community, and vaccination has been recognized as a crucial intervention. To gain insight into public perceptions of COVID-19 vaccines, survey studies and the analysis of social media platforms have been conducted. However, existing methods lack consideration of individual vaccination intentions or status and the relationship between public perceptions and actual vaccine uptake. To address these limitations, this study proposes a text classification approach to identify tweets indicating a user's intent or status on vaccination. A comparative analysis between the proportions of tweets from different categories and real-world vaccination data reveals notable alignment, suggesting that tweets may serve as a precursor to actual vaccination status. Further, regression analysis and time series forecasting were performed to explore the potential of tweet data, demonstrating the significance of incorporating tweet data in predicting future vaccination status. Finally, clustering was applied to the tweet sets with positive and negative labels to gain insights into underlying focuses of each stance.

CCS CONCEPTS

• **Information systems** → **Clustering and classification**; • **Computing methodologies** → **Information extraction**; • **Human-centered computing** → *Social media.*

KEYWORDS

COVID-19, vaccination intent, text classification, tweet analysis, vaccination rate forecasting

ACM Reference Format:
Xintian Li and Aron Culotta. 2023. Forecasting COVID-19 Vaccination Rates using Social Media Data. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30-May 4, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543873.3587639

1 INTRODUCTION

The Coronavirus (COVID-19) pandemic has had a far-reaching and enduring impact on the global community over the past few years, resulting in millions of cases of infection and death and leading to a significant socio-economic crisis [5]. As a means to curb the progression of the pandemic, vaccination has been recognized as a crucial intervention. The success of the vaccination effort is contingent not only on the efficacy and safety of the vaccine, but also on the level of acceptance among the population. To gain insight into COVID-19 vaccination acceptance rates and the factors that influence them, survey studies have been conducted extensively across a variety of countries and regions [18]. However, these studies are often costly and may not be able to keep pace with the dynamic changes as the pandemic and the vaccination process evolve. Given that a substantial number of individuals are now inclined to share their views and activities on social media platforms, these platforms provide a wealth of information regarding individual behaviors and attitudes towards vaccination.

A multitude of studies have been conducted in recent times to understand public perceptions of COVID-19 vaccines through the analysis of Twitter data [6, 7, 11, 17, 20]. While these studies provide valuable insights into public perceptions, they also exhibit several limitations. Firstly, they lack the consideration of individual vaccination intentions or status in the analysis of tweets, which is a crucial aspect of determining the actual level of vaccine uptake and the extent to which public perception influences vaccination behavior. Secondly, there is limited information available on the relationship between public perceptions and the actual trend in vaccine uptake, which is a critical factor in ensuring the effective distribution and coverage of vaccines.

To mitigate the limitations of existing methods, we propose a text classification approach for the identification of tweets that reveal a user's intent or status on vaccination. Our approach involves the geolocation of tweets related to vaccination after the widespread vaccine rollout in December 2020, with a focus on tweets originating within the United States. A total of 1,600 tweets were annotated into four categories based on their vaccination intent or status (vaccinated, positive, neutral, or negative), and a text classifier was trained on this data, achieving an AUC of 0.81. Using the US tweets labeled by our classifier, we conducted a comparative analysis between the proportions of tweets from different identified classes and real-world vaccination data, revealing notable alignment (up to a correlation of 0.84) between data trends. This suggests that tweets related to vaccination may serve as a precursor to the actual vaccination status. Moreover, regression analysis was performed using state-level tweet and vaccination data, demonstrating a moderate association between the proportions of tweets and future vaccination rates. To further explore the potential of tweet data, we also proposed a time series forecasting model for predicting future vaccination status, and our results indicate that incorporating tweet data significantly reduces forecasting error. Additionally, we applied a simple clustering method to the tweet sets with positive and negative labels to gain insights into the underlying focuses of each stance.

WWW '23 Companion, April 30-May 4, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587639

2 RELATED WORK

As the Internet has emerged as a prevalent source of health information, individuals frequently resort to obtaining vaccine information from social media platforms [8]. In addition to acquiring information on vaccines, individuals often use social media to seek social support and engage in conversations with their peers [24]. Hence, social media can be viewed as a crucial channel through which researchers can gain insights into public perceptions of vaccines and public health practitioners can disseminate accurate information to facilitate informed decision-making regarding vaccination.

Studies have revealed that social media has a substantial impact on public attitudes towards vaccines, particularly in contributing to vaccine hesitancy [15, 23]. Given the role played by online misinformation in fostering vaccine hesitancy [10], researchers are making sustained efforts to understand its influence. For instance, Muric et al. [12] constructed a dataset concerning anti-vaccine content and COVID-19 vaccine misinformation by analyzing historical tweets from Twitter accounts that posted tweets containing anti-vaccine keywords. Pierri et al. [14] evaluated the impact of online misinformation on U.S. COVID-19 vaccinations by determining the prevalence of misinformation based on geolocated tweets referencing low-credibility sources, and comparing the results with vaccine survey and uptake data. Additionally, Sharma et al. [19] investigated the characteristics of misinformation and conspiracy groups by identifying suspicious coordinated accounts in Twitter data collected on COVID-19 vaccines.

Gaining an understanding of public perceptions of vaccines is essential for devising effective strategies to influence vaccine decision-making. To that end, many researchers have sought to capture public perceptions from various angles by using social media data. For instance, Saleh et al. [17] analyzed 2.4 million English tweets related to the COVID vaccine during its development, using sentiment and emotion analysis, demographic inference, and topic modeling to examine the evolution of public perception. Huangfu et al. [7] adopted a similar sentiment-based topic modeling approach to study COVID-19 vaccine tweets following vaccine rollout. Lappeman et al. [9] investigated tweets expressing negative sentiment towards COVID-19 vaccines in the U.S. and U.K. to reveal the underlying themes. Luo et al. [11] explored public perceptions of the COVID-19 vaccine by identifying prominent discussion topics on social media platforms using semantic network analysis. Shi et al. [20] compared the psycho-linguistic features of anti-vaxxers on Twitter with those of pro-vaxxers, with the two competing groups being identified by confirming the top users in each community cluster detected. Di Giovanni et al. [6] constructed a tweet dataset that was semi-automatically labeled based on selected hashtags, and then trained a binary classifier to predict the stance of tweets towards vaccines. Zhou and Li [25] devised a framework utilizing autoregressive models to forecast vaccination uptake rates, which draws on both conventional clinical data and innovative web search queries gleaned from Google Trends.

The identification of attitudes towards vaccines in online posts has primarily relied on lexicon and rule-based sentiment analysis tools. However, these methods do not effectively capture the individual's intent or stance on vaccination. While some studies have conducted temporal analysis of sentiment polarity towards COVID-19 vaccines, there is a lack of understanding about the relationship between these discourses and actual vaccination trends. In this context, this paper contributes to the field by presenting a methodology for identifying tweets expressing vaccine intent, conducting temporal and regression analysis to understand factors impacting vaccine decisions, and proposing a forecasting model to predict future vaccination status.

3 DATA

3.1 Twitter data

In order to gain insight into public perceptions of COVID-19 vaccines in the US following the vaccine roll-out, we utilized Twitter content as our primary source of information. To collect tweets related to individual vaccination intent, we employed the full-archive search endpoint provided by the Twitter API [22] to search for all public tweets from December 19, 2020 to August 6, 2021 containing specific keywords, including *"vaccine", "vaccinated", "second shot", "my shot"*, and *"vaxxed"*. The choice of these keywords was predicated on their ability to capture vaccination intent, notwithstanding the possibility of missing out on negative aspects such as misinformation or conspiracy theories that may also influence public perception. To ensure that the tweets reflected users' own opinions and thoughts, we filtered out retweets, tweets containing URLs, and only retained tweets written in English for consistency in processing. The initial retrieval resulted in 26.9 million tweets from across the world, of which only 2% were accompanied by real-time location information.

In order to focus on tweets originating from the US, we utilized the Carmen library [16] to determine the location of each user based on the geo-coordinates provided in the tweets, as well as the location field in the user's Twitter profile. As a result, 11.1 million tweets were successfully geolocated, of which 6.4 million were from the US and were posted by 1.4 million unique users during the specified time period.

3.1.1 Annotation. Based on a preliminary examination of the collected data, four categories were identified to classify tweets according to the expressed vaccination intent. The categories include: 1) *vaccinated*, indicating that the user has received or will soon receive a COVID-19 vaccine; 2) *positive*, conveying support or a favorable view of the vaccine or vaccination without mentioning the user's own status; 3) *negative*, reflecting distrust and concerns about the vaccine; 4) *neutral*, lacking a clear indication of the user's inclination towards the vaccine. A sample of 1,600 tweets was randomly selected from all US tweets and annotated for these categories. The quadratic weighted kappa between the annotations made by two annotators was found to be 0.795 for 100 of these tweets. The annotation resulted in 351 tweets classified as vaccinated, 631 as positive, 280 as negative, and 338 as neutral in terms of vaccination intent. These labeled tweets were then used to train a classifier for categorizing the remaining unlabeled data.

3.2 Vaccination and census data

We utilize the vaccination trend data in the United States provided by the Centers for Disease Control and Prevention (CDC) [4] to compare with our tweet data. This data contains daily updated

national and jurisdictional level statistics, such as the total number of administered doses by date of administration, the percent of the population with at least one dose, and so forth. Given that demographic characteristics play a significant role in determining vaccine uptake across different locations [21], we also use several state-level census data as supplementary information for our tweet data. These data include: 1) uninsured rate: the percent of the population without health insurance [3]; 2) rate 65+: the rate of the population over 65 [1]; 3) poverty rate: the percent of the population below poverty level [2]; and 4) non-metro score: the mean of Rural-Urban Continuum Codes [13] of all counties in a state divided by 10.

4 METHODS AND RESULTS

4.1 Classification

With the limited annotated data set (§3.1.1), our objective is to train a classifier and apply it to the remaining unlabeled tweets, in order to analyze the temporal trends in individuals' vaccination intentions. To this end, we compare two classification approaches: *Logistic Regression* and *Bidirectional Encoder Representations from Transformers* (**BERT**). For Logistic Regression, we experiment with two input feature settings: binary (**LogReg**) and tf-idf (**Tf-idf+LogReg**) features, both using unigrams and bigrams. In the tokenization process, mentions, hashtags, and emojis are identified and transformed into distinct tokens. Additionally, we incorporate simple negation features to capture polarity (e.g., the phrase "not getting vaccinated" becomes "NOT_getting, NOT_vaccinated").

For the **BERT** model, we utilized a pre-trained uncased English language model, which was fine-tuned with our annotated data to perform the sequence classification task. The model architecture comprised of 12 hidden layers with a size of 768 and 12 attention heads, resulting in a total of 110 million parameters. In the tokenization process, we included 181 additional tokens to accommodate for the representation of emojis in the vocabulary, which originally consisted of 30,522 tokens. The model was fine-tuned by training it for 3 epochs with a learning rate of 5×10^{-5}.

Table 1 presents a summary of the accuracy of the models, which were evaluated using 10-fold cross-validation. The precision, recall, and F1 scores represent the weighted average across the four classes identified, with the exception of the 3-class **LogReg** model, where the vaccinated and positive tweets were consolidated into a single class. Additionally, we present the class-specific accuracy of the **Tf-idf+LogReg** model towards the end. It is evident that the model performs better for the vaccinated and positive classes as compared to the neutral and negative classes.

Due to the limitation of the available labeled data, it was challenging to enhance the overall accuracy further. As a result, the 4-class **Tf-idf+LogReg** regression model was selected as the final classifier on account of exhibiting the highest accuracy while also being the easiest to interpret. Upon applying the classifier to the unlabeled data, we obtained the following class distribution: 3.1 million positive tweets, 1.3 million vaccinated tweets, 1.1 million neutral tweets, and 0.9 million negative tweets.

Table 1: Cross-validation accuracy for the classification task regarding vaccination intent.

method	precision	recall	f1	acc	auc
LogReg(3-class)	0.62	0.64	0.63	0.64	0.75
LogReg	0.56	0.57	0.56	0.57	0.79
BERT	0.58	0.59	0.58	0.59	0.80
Tf-idf+LogReg	0.58	0.60	0.58	0.60	0.81
- negative	0.51	0.37	0.43	-	0.80
- neutral	0.51	0.29	0.37	-	0.75
- positive	0.58	0.75	0.66	-	0.77
- vaccinated	0.72	0.79	0.75	-	0.94

4.2 Temporal Analysis

Drawing on the results of the classification procedure described in §4.1, we present the trend in tweets in relation to the national vaccination data over time. The tweets in each class were grouped on a daily basis with respect to Eastern Standard Time (EST) and a 7-day rolling average was calculated to account for daily fluctuations, i.e. the mean value of the previous 7 days as of that day.

An examination of the tweet trend (Figure 1) reveals three distinct peaks, particularly within the positive tweet class: 1) the initial distribution and administration of COVID-19 vaccines in December 2020, 2) the confirmation of the first case of the Delta variant in the US in May 2021, and 3) the rapid growth of both daily test cases and positive cases starting in July 2021, following a sustained period of declining trends. The evolution of the vaccination trend exhibited similar changes to the trends observed in tweets at each of the three stages. Furthermore, there appears to be a correspondence between fluctuations in the vaccination curve, marked by a valley in late February and a peak in April, and similar fluctuations in the tweet curves. To quantify the relationship between the tweet and vaccination trends, we calculated the Pearson correlation coefficient (PCC) between each pair of curves. The correlation between the volume of tweets expressing that the user has been vaccinated has a correlation of .773 with the total number of administered doses; .641 with the number of first dose vaccination; and -.413 with the number of new cases of COVID-19. Thus, the results indicate a strong correlation between the vaccinated tweet trend and the actual vaccination trends.

While Figure 1 depicts the absolute number of tweets in various categories, Figure 2 aims to provide a more nuanced perspective by presenting the proportion of tweets across categories, thereby eliminating the influence of the overall number of tweets on the analysis. The tweet proportion curves demonstrate a strong alignment with the overarching trend as well as with certain fluctuations in the vaccination curve. Our analysis reveals a robust positive correlation between the proportions of tweets categorized as "vaccinated" and the daily number of doses administered, as indicated by a PCC of 0.843. A corresponding decrease in positive tweet proportion is observed as the vaccinated tweet proportion increases, resulting in negative correlations with the vaccination trend (PCC -0.779). Additionally, the neutral and negative tweet proportions also demonstrate negative correlations, as indicated by the PCC values.

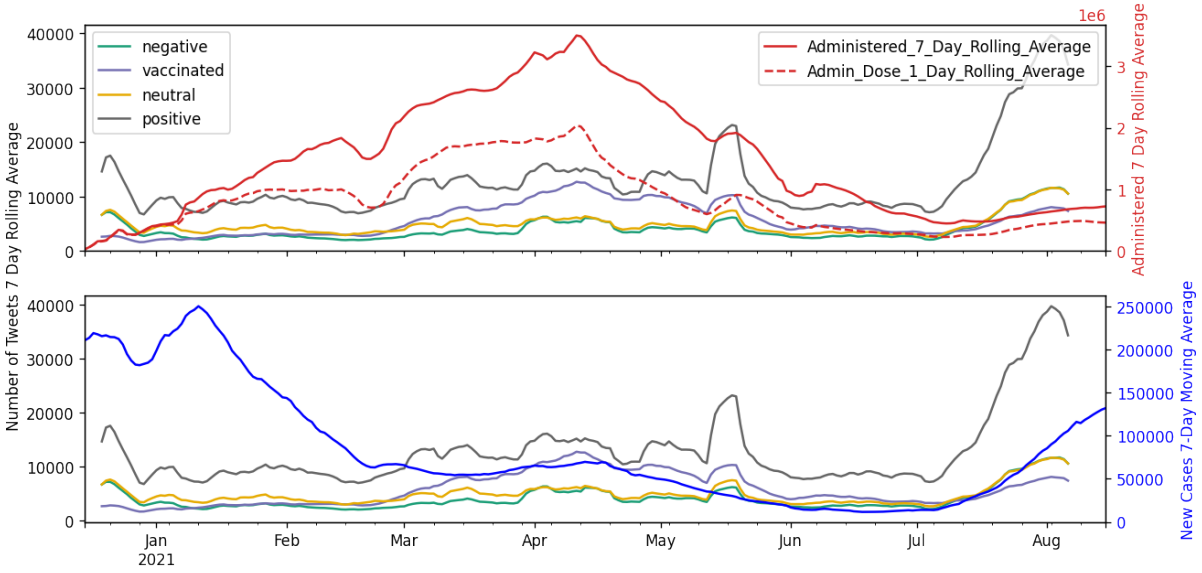

Figure 1: A comparison of the daily tweet trend by class with the vaccination trend (red) and the trend in COVID-19 cases (blue) at a national level. The vaccination trend encompasses both 1) daily administered doses and 2) the number of individuals receiving their first dose of the vaccine. All of the curves in the figure are based on a 7-day moving average calculation. The Pearson correlation coefficients between the number of vaccinated tweets and a) administered doses, b) first dose vaccinations, and c) new cases of COVID-19 are 0.773, 0.641, and -0.413, respectively.

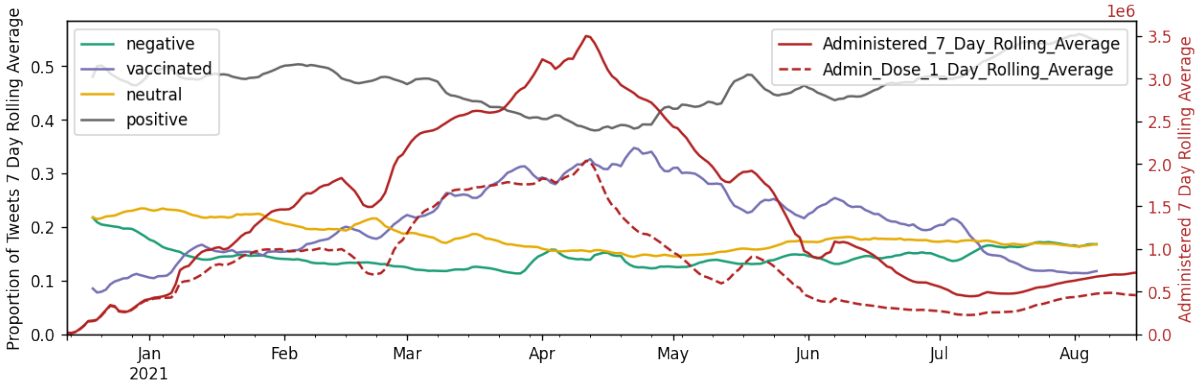

Figure 2: A comparison of the national tweet trend in terms of the proportions of tweets from different classes with the trend in vaccination. The Pearson correlation coefficients between the proportions of each tweet class and the daily administered doses are presented as follows: a) negative, -0.655; b) vaccinated, 0.843; c) neutral, -0.502; d) positive, -0.779.

4.3 Simple Regression Analysis

As an initial analysis to assess whether Twitter trends can serve as a leading indicator of future vaccination activity in the United States, we conducted a simple linear regression analysis. The response variable was the vaccination rate on a specified future date, referred to as the *"target date"*, and the sample was drawn from state-level jurisdictions, including all 50 states and Washington D.C. The explanatory variables can be categorized into three groups. For the census data (3.2), samples were generated for each variable of interest (e.g., poverty rate) in a straightforward manner. In the case of tweet data, a *"source date"* was selected prior to the target date and all classified tweets posted on or before that date were utilized to determine the proportion of tweets classified as a specific category (e.g., "vaccinated"). The vaccination rate at the source date was also considered as an explanatory variable (e.g., "2021-02-28_vr"" indicates the vaccination rate of a state as of February 28th, 2021). Ordinary least-squares models were applied to each explanatory variable based on samples collected from all states. The results of the two date pair settings are depicted in Figure 3, with

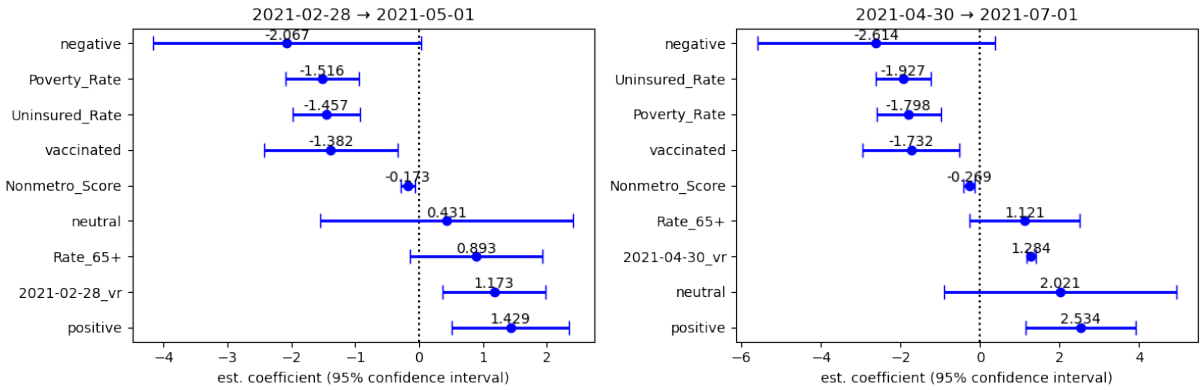

Figure 3: Estimated coefficients of simple linear regressions for the vaccination rate at the target date. The figure displays the results of two regression models: 1) source date: February 28, target date: May 1 (left subfigure); 2) source date: April 30, target date: July 1 (right subfigure). The explanatory variables include: a) census features, b) tweet proportions as of the source date, and c) vaccination rate at the source date.

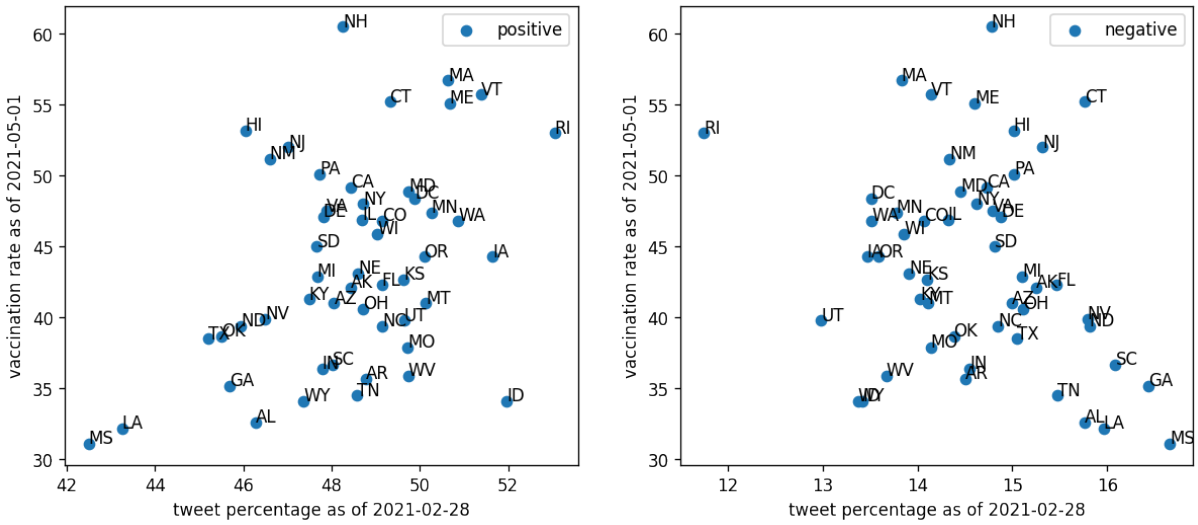

Figure 4: Relationship between the vaccination rates among states as of May 1st, 2021 and the percentage of positive(PCC 0.411)/negative(PCC -0.273) tweets on Twitter as of the end of February, 2021.

the estimated coefficients and 95% confidence intervals presented for comparison.

In both the analyzed settings, we found that the proportion of positive tweets on the source date had the strongest correlation with the future vaccination rate on the target date among different states. Conversely, the proportion of negative tweets demonstrated a contrary influence with a larger variance. This trend can be further observed through the plotting of the vaccination rates of all states on the target date against each of the tweet proportions Figure 4). Additionally, we found that the proportion of neutral tweets was positively correlated with higher vaccination rates among states with large deviation, while the proportion of vaccinated tweets appeared to be a negative indicator in predicting the vaccination rates.

4.4 Forecasting

To more rigorously assess the utility of the Twitter data for building real-time estimates of future vaccination activity, we next present a vaccine administration forecasting model based on a single-step time series forecasting approach. Our model utilizes aggregated information within a designated time window to predict the number of vaccine doses administered in a given region within a specified time range in the near future (as depicted in Figure 6). The time window is comprised of multiple consecutive time steps of uniform length (e.g., one week). In a given time step t_i, the known information, such as tweet trend and census data, is combined to form a feature vector $\mathbf{x}_i$, while the number of administered doses, normalized by the population of the corresponding region, is represented as y_i. We aim to fit a model $\mathcal{F}_{s,g,w}$ that approximates the

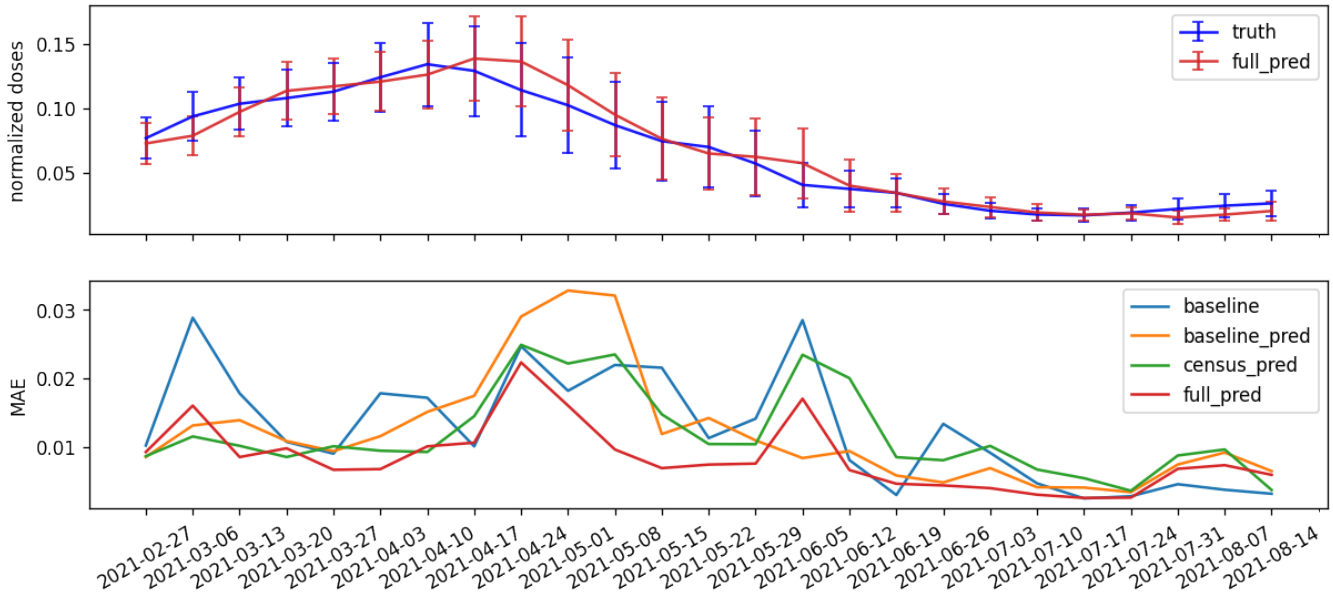

Figure 5: Forecasting on normalized doses using ARLSTM model ($s = 6, g = 0, w = 2$).

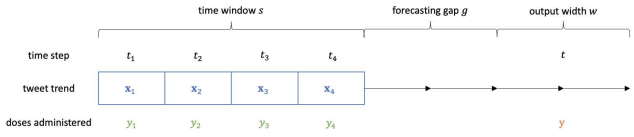

Figure 6: A time-series sample in the vaccine administration forecasting model.

relationship between the normalized number of administered doses in a future time range t, denoted by y, and the aggregated information in a specified time window, represented by a set of feature pairs $(\mathbf{x}_i, y_i)$ for each time step. The width of the time window, represented by s, is the number of time steps it encompasses. The forecasting gap, represented by g, is the number of time steps between the final time step t_s and the target time t. The width of the target time range t is denoted by w and is expressed in number of time steps. To generate training data, we consider input instances of the model $\mathbf{z}_j = \{(\mathbf{x}_i, y_i)\}_{i=j}^{s+1-j}$ and their corresponding output values y_j. We can generate n such pairs $\{(\mathbf{z}_j, \mathrm{y}_j)\}_{j=1}^{n}$ for a region by incrementally shifting the time window forward one step at a time, starting with an initial example.

4.4.1 Input features. As previously mentioned, at each time step t_i, the input consists of two components, $\mathbf{x}_i$ and y_i, where y_i represents the quantity of doses administered. To assess the efficacy of tweet-related features, three distinct configurations of $\mathbf{x}_i$ have been formulated. In the first configuration, referred to as the baseline setting (**baseline pred**), only y_i is utilized as the input for each time step, and $\mathbf{x}_i$ is excluded. In the second configuration (**census pred**), census data is incorporated into the input, such that $\mathbf{x}_i$ encompasses several census features of the region in question. It should be noted that the census features remain constant across different time steps.

In the third configuration, referred to as the complete setting (**full pred**), the tweet features are added to $\mathbf{x}_i$. These features are obtained through a procedure that is similar to the one described in §4.3. For a specific region and time step, the cumulative number of tweets in each class is tallied from an initial date to the end date of that time step. Subsequently, the proportions of tweets among all four classes are calculated and employed as feature values. This means that the resulting features not only depict the situation within the given time step, but also encompass historical information pertaining to the region in question.

4.4.2 Experiments and results. We implement a time-rolling approach for the model's training and testing, which commences on December 19, 2020 with a time step length of one week. Given the consideration of 51 state-level jurisdictions, a training example can be generated for each time window from each jurisdiction. As the time window advances by one step, an additional 51 examples are added to the training data set, and the model is refitted using the augmented training data. The testing is carried out concurrently by utilizing the trained model on examples from the subsequent time window and calculating the errors between the obtained results and the actual values.

In recognition of the varying scales of the input features, standardization is performed prior to training and testing by normalizing the features using the standard score. To be more specific, the mean and standard deviation of each feature are calculated based on all time steps included in the current training data, taking into account that each time step is only counted once, even if it appears in multiple successive time windows.

In our efforts to evaluate the efficacy of forecasting, we test a range of models, including linear regression (**LinearReg**), Ridge

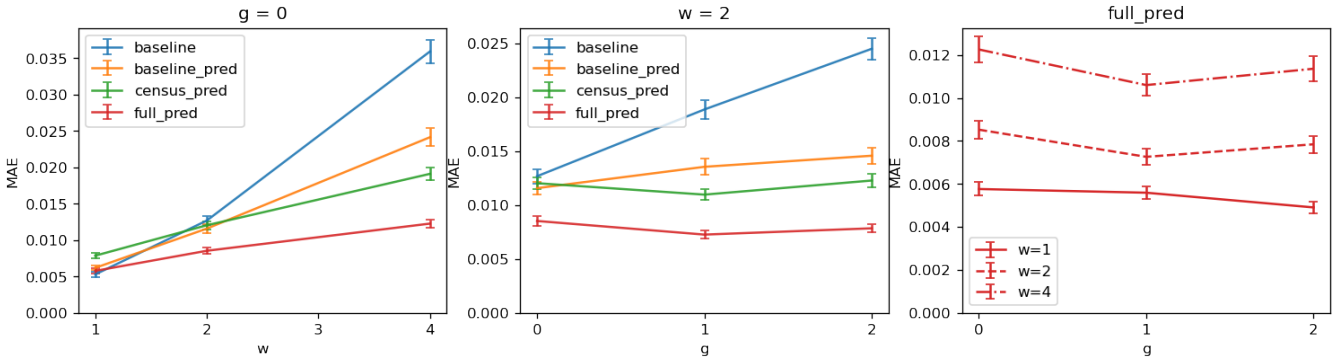

Figure 7: Performance of ARLSTM for various forecasting settings ($s = 6$), where s is the width of the time window in weeks, g is the forecasting gap in weeks, and w is the width of the target time range in weeks.

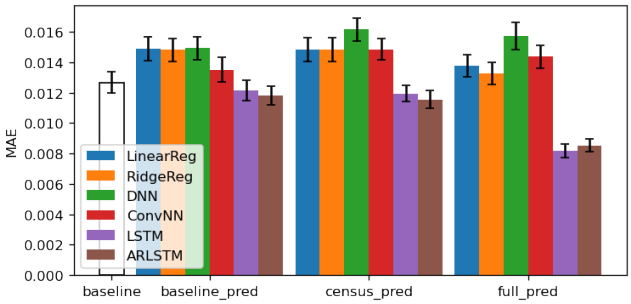

Figure 8: Performance of different forecasting models measured by MAEs with 95% bootstrap confidence interval across all states and queried time ranges ($s = 6, w = 2, g = 0$).

regression (**RidgeReg**), deep neural network (**DNN**), convolutional neural network (**ConvNN**), Long short-term memory (**LSTM**), and auto-regressive LSTM (**ARLSTM**). For the initial three models, each input example ($\mathbf{z}_j$) is first transformed into a feature vector before being fit to the model. The DNN model consists of two hidden layers, each containing 64 units. In the ConvNN model, the inputs are first passed through a convolutional layer with 32 filters, where each kernel has a size of 3, before being flattened and fed into a hidden layer of 32 units. The LSTM model comprises an LSTM layer of 32 units and a 32-unit hidden layer. In the ARLSTM model, $(g + w)$ LSTM steps are executed to produce the final prediction, whereas only one step is performed in the LSTM model. The mean squared error is employed as the loss function for the four neural network models, which are trained for 40 epochs using the Adam optimizer with a learning rate of 0.005.

The performance of the models is evaluated using mean absolute errors (MAEs) under different forecasting and feature settings. In order to reduce the impact of randomness in the results, the training and testing processes for the neural network models are repeated three times and the average predictions are used. Figure 5 illustrates the testing results of the **ARLSTM** model for the case when $s = 6$, $g = 0$, and $w = 2$. The upper plot shows the mean true/predicted normalized administered doses for all states in the queried time range, with the error bars indicating the standard deviation among the states. The lower plot reports the prediction errors over time for different input settings. The results indicate that the **full pred** setting generally yields the best predictions, especially when the actual vaccination trend is decreasing. The **baseline** estimates are calculated as $\hat{y}_{j,\text{baseline}} = \mathcal{F}^{\text{baseline}}_{s,g,w}(\mathbf{z}_j) = w \cdot \mathbf{x}_{s+1-j}$, where the normalized doses of the last time step in the input time window are simply multiplied by the width of the queried time range.

Figure 8 shows the comparison of different models under various feature settings using the same forecasting parameter setting. The empirical bootstrap is applied to calculate 500 MAEs and the 95% confidence interval is determined for each group of predictions under the same model and input settings. The results indicate that only the **LSTM** and the **ARLSTM** models outperform the baseline estimate for all feature settings. The input setting with tweet features (**full pred**) significantly improves the performance of LSTM and ARLSTM, demonstrating the benefits of incorporating tweet information in the forecasting procedure.

Figure 7 shows the evaluation results of **ARLSTM** under different forecasting parameter settings with 95% bootstrap confidence intervals. The results suggest that the MAE increases as the forecasting time span (w) expands and the error of **full pred** rises the slowest. However, when the forecasting gap (g) is fixed to 0 and the forecasting time span is one week, the forecasting model does not outperform the baseline estimates, indicating that it is difficult to make accurate predictions within a short time range when the normalized doses of different states differ greatly. As the forecasting time span increases, the variance across different states decreases, enabling the forecasting model to effectively predict the future values. Additionally, the results also demonstrate that the tweet-related features (**full pred**) consistently improve the accuracy of the forecasts under most settings.

4.5 Clustering

To gain insights into the key factors affecting people's intent towards COVID vaccines, we conducted a clustering analysis of the classified tweets. The clustering was performed using a K-means algorithm, where each tweet was represented as a vector of term frequencies and normalized using the L2 norm. For this analysis, we

sampled 50,000 tweets from both the positive and negative classes from the classified tweets.

The tokenization process was similar to that used in the classification task and involved removing stop words and frequently occurring terms. The resulting clusters were characterized by their size and center vector, and the top terms were identified by sorting the coefficients of the center vector. To determine the best number of clusters for the K-means algorithm, several settings were tested, with the aim of obtaining meaningful top terms for the majority of clusters and as many clusters as possible. The smallest cluster was also required to have a sufficient number of tweets. With these considerations, the number of clusters was set to 32 for both classes.

The results of the clustering analysis are presented in Table 2 in the Appendix, which lists the clusters in descending order of size and shows the top 4 terms for each cluster. For both the positive and negative classes, the largest cluster was omitted as it was sparse due to its extremely large size.

The examination of positive tweets sheds light on several prevalent themes that are frequently mentioned by individuals who express their support for or hold positive views towards vaccination. These themes are a reflection of the general sentiment of individuals who support vaccination, and highlight the key reasons behind their positive outlook. Some of the most common expressions include "ending the pandemic", which underscores the importance of vaccination in bringing an end to the ongoing COVID-19 pandemic. "Countering anti-vaxxers" is another common expression that highlights the need for individuals to stand up against misinformation and negative propaganda surrounding vaccination. "Life-saving" is a testament to the crucial role that vaccination plays in saving lives, while "feeling better" reflects the improved health and well-being that individuals experience after being vaccinated.

On the other hand, the analysis of negative tweets provides insights into the key concerns that individuals have regarding vaccination. These concerns are reflected in the presence of terms such as "experimental", which highlights the uncertainty surrounding the long-term effects of the COVID vaccines. "Passport" is another term that is frequently mentioned, and reflects the worries that individuals have regarding the potential use of vaccination as a form of coercion or discrimination. "Long-term effects", "FDA approved", and "blood clots" are expressions that reflect the general unease that individuals have regarding the safety and efficacy of the vaccines.

It is worth noting that certain terms, such as "mask", "death", "school", and "immunity", are present in both positive and negative tweets. However, the contexts in which these terms are used differ, with terms such as "herd immunity" appearing in positive tweets, while "natural immunity" is found in negative tweets. This highlights the importance of considering the wider context in which these terms are used, and underscores the need for a nuanced and sophisticated analysis of the data.

5 CONCLUSION

In this paper, we present an approach for text classification to identify tweets related to COVID-19 vaccination status or intent. The results of our subsequent temporal and regression analysis reveal strong correlations between tweet proportions of different classes and the actual vaccination trend. Furthermore, our forecasting model shows that tweet-related features significantly enhance the accuracy of state-level vaccination forecasts, suggesting that tweet trends may serve as a useful precursor of the actual vaccination status. Finally, our clustering analysis uncovers recurring themes and key concerns among individuals regarding vaccination.

It is important to acknowledge the limitations of our study, as is common with many studies based on social media data. First, it is necessary to recognize that Twitter users may not be a representative sample of the general population, which can limit the generalizability of our findings, particularly for specific states due to the large variation in the number of tweets among them. In addition, our study's analysis is restricted to the English language, potentially excluding data from non-English speaking populations. Furthermore, our text classification approach relied on a classifier trained on data sampled through select keywords, and not validated on all Twitter messages. Additionally, the classifier's accuracy is imperfect, which may influence the subsequent analysis and interpretation of the data. Finally, it is crucial to consider the limitations of self-reporting and the potential for disparities between attitudes expressed on social media and those held in real life.

In future work, a promising direction is to broaden the scope of data sources beyond Twitter to include other social media platforms. This could provide a more comprehensive understanding of public attitudes related to vaccination and address the potential biases of a single platform. Additionally, incorporating region-specific data, such as public health policies, may provide a more nuanced understanding of vaccination behaviors and allow for the identification of reasons for variations in vaccination rates between different states. Moreover, exploring domain adaptation techniques may improve the accuracy of vaccination status classification considering the evolving vaccination practices and policies. Overall, these future directions may enhance our comprehension of public attitudes and behaviors towards vaccination and lead to more effective vaccination campaigns.

ACKNOWLEDGMENTS

This research is supported in part by the National Science Foundation under grant #2133960 and by the Harold L. and Heather E. Jurist Center of Excellence for Artificial Intelligence at Tulane University.

REFERENCES

[1] ACS. 2019. Age and Sex. Retrieved Nov 22, 2021 from https://data.census.gov/cedsci/table?t=Age%20and%20Sex&g=0100000US%240400000&tid=ACSST5Y2019.S0101&hidePreview=true&moe=false

[2] ACS. 2019. Poverty Status in the Past 12 Months. Retrieved Nov 22, 2021 from https://data.census.gov/cedsci/table?g=0100000US%240400000&tid=ACSST5Y2019.S1701&hidePreview=true&moe=false

[3] ACS. 2019. Selected Characteristics of Health Insurance Coverage in the United States. Retrieved Nov 22, 2021 from https://data.census.gov/cedsci/table?t=Health%20Insurance&g=0100000US%240400000&tid=ACSST5Y2019.S2701&hidePreview=true&moe=false

[4] CDC. 2021. COVID-19 Vaccination Trends in the United States, National and Jurisdictional. Retrieved Nov 22, 2021 from https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2

[5] Vicente Javier Clemente-Suárez, Eduardo Navarro-Jiménez, Libertad Moreno-Luna, María Concepción Saavedra-Serrano, Manuel Jimenez, Juan Antonio Simón, and Jose Francisco Tornero-Aguilera. 2021. The impact of the COVID-19 pandemic on social, health, and economy. *Sustainability* 13, 11 (2021), 6314.

[6] Marco Di Giovanni, Lorenzo Corti, Silvio Pavanetto, Francesco Pierri, Andrea Tocchetti, and Marco Giovanni Brambilla. 2021. A Content-based Approach for the Analysis and Classification of Vaccine-related Stances on Twitter: the Italian Scenario. In *Information Credibility and Alternative Realities in Troubled Democracies@ ICWSM 2021*. 1–6.
[7] Luwen Huangfu, Yiwen Mo, Peijie Zhang, Daniel Dajun Zeng, Saike He, et al. 2022. COVID-19 Vaccine Tweets After Vaccine Rollout: Sentiment–Based Topic Modeling. *Journal of medical Internet research* 24, 2 (2022), e31726.
[8] Seth C Kalichman and Christopher Kegler. 2015. Vaccine-related internet search activity predicts H1N1 and HPV vaccine coverage: Implications for vaccine acceptance. *Journal of health communication* 20, 3 (2015), 259–265.
[9] James Lappeman, Keneilwe Munyai, and Benjamin Mugo Kagina. 2021. Negative sentiment towards COVID-19 vaccines: A comparative study of USA and UK social media posts before vaccination rollout. *F1000Research* 10, 472 (2021), 472.
[10] Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour* 5, 3 (2021), 337–348.
[11] Chen Luo, Anfan Chen, Botao Cui, and Wang Liao. 2021. Exploring public perceptions of the COVID-19 vaccine online from a cultural perspective: Semantic network analysis of two social media platforms in the United States and China. *Telematics and Informatics* 65 (2021), 101712.
[12] Goran Muric, Yusong Wu, Emilio Ferrara, et al. 2021. COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies. *JMIR public health and surveillance* 7, 11 (2021), e30642.
[13] Timothy Parker. 2013. Rural-Urban Continuum Codes. (2013).
[14] Francesco Pierri, Brea Perry, Matthew R DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. 2021. The impact of online misinformation on US COVID-19 vaccinations. *arXiv preprint arXiv:2104.10635* (2021).
[15] Neha Puri, Eric A Coomes, Hourmazd Haghbayan, and Keith Gunaratne. 2020. Social media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious diseases. *Human vaccines & immunotherapeutics* 16, 11 (2020), 2586–2593.
[16] Roger Que and Mark Dredze. 2017. Carmen: a library for geolocating tweets. Retrieved May 25, 2021 from https://github.com/mdredze/carmen-python
[17] Sameh Nagui Saleh, Samuel A McDonald, Mujeeb A Basit, Sanat Kumar, Reuben J Arasaratnam, Trish M Perl, Christoph U Lehmann, and Richard J Medford. 2021. Public perception of COVID-19 vaccines through analysis of Twitter content and users. *medRxiv* (2021).
[18] Malik Sallam. 2021. COVID-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. *Vaccines* 9, 2 (2021), 160.
[19] Karishma Sharma, Yizhou Zhang, and Yan Liu. 2021. COVID-19 Vaccine Misinformation Campaigns and Social Media Narratives. *arXiv preprint arXiv:2106.08423* (2021).
[20] Jialiang Shi, Piyush Ghasiya, and Kazutoshi Sasahara. 2021. Psycho-linguistic differences among competing vaccination communities on social media. *arXiv preprint arXiv:2111.05237* (2021).
[21] J Tolbert, K Orgera, R Garfield, J Kates, and S Artiga. 2021. Vaccination is local: COVID-19 vaccination rates vary by county and key characteristics. *KFF* (2021).
[22] Twitter. 2021. Search the full archive of Tweets. Retrieved Nov 22, 2021 from https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all
[23] Steven Lloyd Wilson and Charles Wiysonge. 2020. Social media and vaccine hesitancy. *BMJ Global Health* 5, 10 (2020), e004206.
[24] Yuehua Zhao and Jin Zhang. 2017. Consumer health information seeking in social media: a literature review. *Health Information & Libraries Journal* 34, 4 (2017), 268–283.
[25] Xingzuo Zhou and Yiang Li. 2022. Forecasting the COVID-19 vaccine uptake rate: an infodemiological study in the US. *Human vaccines & immunotherapeutics* 18, 1 (2022), 2017216.

APPENDIX

Table 2 below shows the results of the clustering analysis on positive and negative tweets.

Table 2: Top terms of clusters in positive and negative tweets derived using K-means. For each cluster, the terms are in descending order of importance.

positive		negative	
size	top terms	size	top terms
23206	mask,better,state,help	21971	spread,@potus,biden,live
3635	mask,wear,wear mask,fully	2753	death,death death,hospitalized,cases
1725	death,sick,cases,americans	2082	effect,death,pfizer,second
1203	trump,president,biden,credit	1567	experimental,gene,therapy,gene therapy
1049	risk,high,high risk,death	1534	passport,id,vote,voter
986	fully,mask,cdc,weeks	1443	work,mask,death,mask work
978	safe,effective,stay,safe effective	1264	mask,wear,wear mask,cdc
970	vaccination,proof,rates,mask	1236	long,term,long term,effect
915	home,wait,stay,stay home	1100	risk,death,high,high risk
911	pandemic,end,mask,end pandemic	1044	wait,effect,long,death
899	second,sense,wrong,common	923	sick,death,sick death,work
846	kids,school,parents,mask	845	approved,fda,fda approved,approved fda
838	actually,protect,mask,help	836	school,kids,choice,personal
831	anti,vaxxers,anti vaxxers,vax	833	government,passport,money,death
781	appointment,available,appointment available,#covid9	818	trump,biden,president,death
756	life,saving,life saving,normal	815	actually,blood,clots,blood clots
743	line,love,workers,@marcorubio	771	flu,death,flu flu,shots
739	feel,better,feel better,comfortable	760	better,cdc,fully,cases
710	spread,mask,death,prevent	688	world,understand,china,rest
701	family,friends,members,family members	642	immune,dose,antibodies,immune systems
700	immunity,herd,herd immunity,natural	637	children,school,death,experimental
656	population,fully,million,half	601	trust,government,science,trust government
599	likely,death,spread,likely death	558	body,choice,body choice,abortion
566	effective,preventing,effective preventing,death	545	believe,death,work,effect
548	pfizer,moderna,pfizer moderna,moderna pfizer	525	ones,shots,hospital,loved
535	soon,possible,soon possible,available	513	chance,wrong,reaction,adverse
532	teachers,school,schools,staff	508	test,positive,test positive,negative
463	variant,delta,delta variant,mask	457	safe,effect,death,effective
459	free,donut,krispy,mask	445	forced,experimental,work,choice
435	mom,dad,appointment,week	365	immunity,natural,natural immunity,herd immunity
416	gotten,sick,death,fully	268	mandate,mask,mask mandate,school
316	rate,survival,survival rate,recovery	207	johnson,johnson johnson,blood,clots