COVID-19 Prediction via Vaccine Sentiment Analysis on Twitter

Jinyang Han
Georgia Institute of Technology
Atlanta, Georgia, USA
jhan411@gatech.edu

Jingjing Ye
Georgia Institute of Technology
Atlanta, Georgia, USA
jye312@gatech.edu

Qingqing Wu
Georgia Institute of Technology
Atlanta, Georgia, USA
qwu325@gatech.edu

ABSTRACT

As the COVID-19 pandemic spread globally, vaccines have been expected as the ultimate effective mechanism of defense. Issues related to vaccines receive lots of public attention. In this study, we plan to reveal public opinion towards COVID-19 vaccines with Twitter data and how such sentiment influences vaccination and cases/deaths in the US. Our results will provide insight on vaccine campaign and vaccination plan for future epidemics.

KEYWORDS

sentiment analysis, vaccination prediction, natural language processing, machine learning

ACM Reference Format:
Jinyang Han, Jingjing Ye, and Qingqing Wu. 2018. COVID-19 Prediction via Vaccine Sentiment Analysis on Twitter. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

There is no doubt that in the past two years, COVID-19 has subverted the previous lifestyle. But fortunately, the whole nation is on the path to being normal because of all adult vaccination programs released in March 2021. And the efficacy of Pfizer vaccine is 95% (95% credible interval, 90.3 - 97.6) for preventing COVID-19 after fully vaccinated, and the results of subgroups defined by sex, age, race and etc behave similarly[21].

As for social media platforms, the active users have skyrocketed in past decades because of the popularity of smartphones and comprehensive network coverage. For example, Facebook had 1.91 billion daily active users (DAU) on average for June 2020[11], while the DAU of Twitter increased 20% and reached 199 million in the first quarter of 2021[24].

People are increasingly inclined to use social media to record emotions and opinions, providing unprecedented rich resources for studying sentiment propagation and epidemic transmission [10] [26]. Moreover, during the COVID-19 outbreak period, people have increased the social media platforms usage and dependence[6] to remain connection. Considering the vital of vaccines and soaring social media usage, it will be useful for future vaccine campaigns and future epidemics' policy-making if public sentiment's influence on vaccination is discovered.

In this research, we plan to use the Twitter data from the Panacea Lab[3], combining data of the number of vaccine intake from the CDC to analyze the effect of public opinion towards vaccines and further predict whether they can affect the vaccination and confirmed cases/deaths.

Specifically, the problem can be formulated as: Given Twitter users' sentiment about vaccines, the detailed numbers of vaccination and cases/death, build a model to estimate the impact of public sentiment on vaccine intake and cases/deaths. Moreover, sentiment impact, vaccination and cases/deaths can be modeled for illustration among three large cities.

2 RESPONSE TO COMMENTS ON PROPOSAL

2.1 Reference

2.1.1 Comment. It might be better to also check some top journals (i.e., science, nature, nature human behavior, etc.) and the conference papers.

2.1.2 Response. We summarize and review more papers from top journals, which can be seen in literature review. Besides, we include several reference that are not covered in the proposal.

2.2 Predicting objective

2.2.1 Comment. Be more specific about what the research wants to predict, such as mortality, hospitalization, or cases, etc.

2.2.2 Response. The model will predict vaccination rate and cases/death in US through vaccine sentiment analysis. Further, the cross-sectional comparison will also be implemented to show whether the difference in sentiment is related to difference in vaccination and cases/death across states.

2.3 Expand Research

2.3.1 Comment. Be better to add some controlled case studies. For instance, the statistical hypothesis testing can be performed if increasing vaccine sentiment leads to decreases in cases or mortality in some locations, let say big cities.

2.3.2 Response. We plan to expand our research to estimate the relationship between vaccine sentiment and the future new cases and mortality in three cities: New York, Los Angeles and Houston.

3 LITERATURE REVIEW

3.1 Social Media and Epidemics Forecast

Users from various places upload an abundant amount of raw data in form of text, photos, videos and audio on social media. Information are generated and circulated on social media, which can

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/1122445.1122456

reflect and in turn shape people's thoughts and behaviors. Numerous studies suggest various means to apply these big data sets to the area of public health.

These studies usually focus on the spatio-temporal properties of social media users' sentiment to identify possible disease outbreaks [2] [9]. Using the distribution of total tweet volume, [5] detects a temporal lag of 6–27 days between the rises in the number of COVID-19 related tweets and officially reported deaths in various UK cities.

[23] reveals that the collective wisdom of the crowds at early stages of the pandemic can predict the extent of mortality reflecting the regional severity of the pandemic almost a month later, based on the intensity of initial COVID-19 related tweet attention at the beginning of the pandemic across Italian, Spanish, and United States regions.

3.2 Methods to Analyze Sentiment in Twitter

Sentiment analysis, emotion analysis, topic modeling, and other tools are implemented to explore public sentiment and emotions in Twitter. Lexicon based method, machine learning method or a mix of both methods are usually used in implementing sentiment analysis[9].

Lexicon based method is an unsupervised learning method, which does not require training data and only depends on the dictionary. Words are classified as positive or negative in the polarity lexicons. The occurrences of the terms in the text data are calculated and then transformed into sentiment indicators. This method highly depends on the quality of the lexical resources. The drawbacks are also obvious, such as words can have different meanings based on the context, sentiment words may not express any sentiment. [1][7][17]

Machine learning method is a supervise learning method which requires training data. The most common used method in machine learning method is the SVM and Naive Bayes model. Naive Bayes is successful when applied on well-formed text corpus while support vector machine gives a good performance for low shape dataset ([14][9]). Using training data consists of Twitter messages with emoticons obtained through automated means, [13] shows that machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) have accuracy above 80% when trained with emoticon data.

[8] shows that lexicon and machine learning approaches are similar in accuracy, both achieving higher accuracy when classifying positive sentiment than negative sentiment. The combined approach demonstrates significantly improved performance in classifying positive sentiment.

3.3 Vaccine Sentiment in Twitter

There have been several works analyzing Twitter datasets to reveal vaccine sentiment. [15] analyzes tweets with location information in the US and reveals raising public confidence in vaccines in most states with increasing positive sentiment and decreasing negative sentiment. Besides, critical social/international events (such as clinical trials from Moderna or Pfizer), announcements of political leaders and authorities (such as Donald Trump tweeting "Great News on Vaccines!") and vaccine-adverse conspiracy (such as claim related to Bill Gate that the pandemic is a cover for his plan to implant trackable microchips made by Microsoft) may have potential impacts on public opinion towards COVID-19 vaccines[15][20].

People still take a positive attitude towards vaccination instead of some adversarial effects of some of the vaccines and the emotion of trust regards to vaccine dominates the discussion continually[22] [18].

Further, there exists some geospatial difference in public opinion on vaccines since negative sentiments and emotions are more obvious in some states. [15]

However, limited studies have researched the impact of social media such as Twitter on public vaccination behavior using empirical data. Our research hopes to fill this gap to explore whether and how vaccine sentiment on Twitter influences vaccination rate as well as the epidemic.

4 TECHNICAL METHODOLOGY

4.1 Sentiment Analysis

As illustrated in the literature review, lexicon method and machine learning method are implemented to analyze text sentiment. In this study, three methods are used to examine the fraction of tweets with negative, neutral and positive sentiments related to vaccine.

4.1.1 Naive Bayes. Naive Bayes is a simple model which works well on text categorization[19]. A Naive Bayes model is trained in this study. Naive Bayes model assumes words position does not matter and relies on a very simple representation of document, that is bag of words. Documents are represented by feature and naive Bayes model assumes the feature probabilities are independent given the class.

Class $c*$ is assigned to tweet d, that is

$$\begin{aligned} c* &= arg\max_c P_{NB}(c|d) \\ &= arg\max_c \frac{P(c)(d|c)}{P(d)} \\ &= arg\max_c P(d|c)P(c) \\ &= arg\max_c P(x_1,\ldots,x_m|c)P(c) \\ &= arg\max_c p(c)\prod_i P(x_i|c), \end{aligned}$$

where x_i represents a feature and there are m features. The parameters can be estimated through maximum likelihood estimate and Laplace (add-1) smoothing is utilized for unseen features, that is

$$\begin{aligned} P(x_i|c) &= \frac{count(x_i,c)+1}{\sum_{x_j\in V}(count(x_j,c)+1)} \\ &= \frac{count(x_i,c)+1}{\sum_{x_j\in V} count(x_j,c)+|V|}. \end{aligned}$$

This study implements two Naive Bayes models trained on two different datasets to classify sentiment in tweets.

- TextBlob
  Users can determine the opinion or emotion that a text holds, and the sentiment function of this software offers users a polarity and subjectivity value after analysis. The polarity value ranges from -1 to 1, where -1 indicates it is a negative

statement. TextBlob's default sentiment analysis is trained on customer reviews hand-tagged with values for polarity and subjectivity. Another option in TextBlob is NaiveBayesAnalyzer, which is trained on movie reviews associated with positive or negative rating scores.

- Classifier Trained on the Sentiment140 dataset
Sentiment140[13] dataset is widely used in analysing the sentiment in tweets. The dataset contains about 1.6 million tweets collected through keyword search and annotated automatically by detecting emoticons. Tweets are determined to have positive, neutral, or negative sentiment.
A naive Bayes classifier is trained on this dataset and implemented to classify the sentiment in tweets related to vaccine as either positive or negative.

4.1.2 VADER. VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based feeling analysis instrument that is explicitly sensitive to suppositions communicated in web-based media [16]. It is trained by asking and paying people to score a very big list of words. VADER utilizes a mix of lexical highlights that are, for the most part, marked by their semantic direction as one or the other positive or negative. Thus, VADER not only tells about the Polarity score yet, in addition, it tells us concerning how positive or negative a conclusion is.

4.2 Machine Learning Model

According to the vaccine data we found, supervised deep learning models will be suitable. More recently, machine learning models have drawn attention and have established themselves as serious contenders to classical statistical models in the forecasting community[4]. Models that will be implemented include Random Forest and Multilayer Perceptron.

4.2.1 Random Forest. Random forest is a widely used heuristic machine learning prediction algorithm known to perform well at a variety of predictive tasks by combining a large number of regression or classification trees into an ensemble[25].

4.2.2 Multilayer Perceptron. Assuming adequate data and computing resources, if a strong theoretical understanding of the problem is available, a full numerical model is perhaps the most desirable solution. However, in general, as the complexity of a problem increases, the theoretical understanding decreases (due to ill-defined interactions between systems) and statistical approaches are required. Recently, the use of neural networks, and in particular the multilayer perceptron, has been shown to be effective alternatives to more traditional statistical technique[12]. Here we assume MP can get better effect than random forest.

MLPs are feed forward neural networks which are typically composed of several layers of nodes with unidirectional connections, often trained by back propagation[27]. The learning process of MLP network is based on the data samples composed of the N-dimensional input vector x and the M-dimensional desired output vector d, called destination. By processing the input vector x, the MLP produces the output signal vector $y(x, w)$ where w is the vector of adapted weights. The error signal produced actuates a control mechanism of the learning algorithm. The corrective adjustments are designed to make the output signal $y_k(k = 1, 2, , M)$ to the desired response dk in a step by step manner.

Gradient algorithm to get minimization. Adaptation of weights is performed step by step in gradient algorithm. $p(k$ is the direction of minimization in kth step, is the learning coefficient, and w is the adaptation coefficien

The learning algorithm of MLP is based on the minimization of the error function defined on the learning set (x_i, d_i) for $i = 1, 2, , N$ using the Euclidean norm:

$$E(w) = \frac{1}{2} \sum_{i=1}^{N} \|y\,(x_i, w) - d_i\|^2 .$$

Adaptation of weights is performed step by step

$$w(k+1) = w(k) + \gamma p(k),$$

where $p(k)$ is the direction of minimization in kth step,γ is the learning coefficient, and w is the adaptation coefficient.

Most effective is the Levenberg–Marquard algorithm for medium size networks and conjugate gradient for large size networks.

Levenberg–Marquard algorithm

Least square formulation of learning problem is exploited:

$$E(w) = \frac{1}{2} \sum_{i=1}^{M} (y_i(w) - d_i)^2 .$$

Solved by using second order method of Newton type:

$$p(k) = -G(k)^{-1} g(k),$$

where $g(k) = \frac{\partial E}{\partial w(k)}$ is the gradient of error function Eq. $G(k)$ is the Hessian approximation, determined by applying the Jacobian matrix $J(k)$:

$$G(k) = J(k)^T J(k) + v.$$

In this equation the Jacobian matrix J is equal

$$J = \frac{\partial e}{\partial w} e = [y_i(w) - d_i, \ldots, y_M(w) - d_M]^T .$$

Conjugate gradient

Direction $p(k)$ is evaluated according to the formula.

$$p(k) = -g(k) + \beta p(k-1),$$

where the conjugate coefficient β is usually determined according to the Polak-Ribiere rule:

$$p(k) = \frac{g(k)^T (g(k) - g(k-1))}{g(k-1)^T g(k-1)}.$$

5 DATA COLLECTION AND INITIAL FINDINGS

As mentioned in *Introduction*, four datasets will be utilized in our research.

5.1 Tweets Data from Panacea Lab

5.1.1 Data Collection. Tweets dataset from Panacea Lab includes tweets with keywords containing "COVID-19" or "vaccine" from March 22, 2020 until now. Hydrating all tweets are very time-consuming, thus we pick a time period to show some initial results.

5.1.2 Data Description. The daily data set contains two *tsv* files. Both include three columns, "tweet_id", "date" and "time", which represent the tweet id of the collected post, posting date and posting time respectively. The difference is that one includes all original tweets and re-tweets with the keywords, the another file filters out all the re-tweets. In this work, we choose the clean file (without re-tweets) for research.

Additionally, content, language as well as location are all stored stratified under metadata "tweet_id". These information is most relevant to our research since we plan to extract sentiment information using content analysis. "Hydrator" will be used to recover the full tweet content and geolocation data based on the tweet IDs.

Specifically, we screened out the English-only posting tweets and randomly pick 10% of the IDs to form the twitter dataset. A Number of daily final selected IDs is plotted in Figure 1.

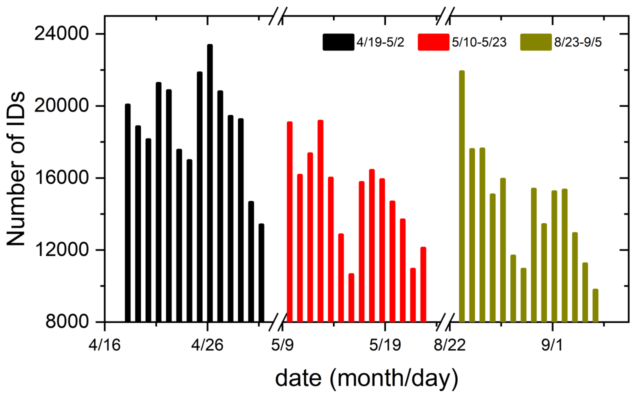

Figure 1: Number of Final Selected IDs On Each Day

5.1.3 Initial Findings on Vaccine Sentiment. Figure 2 shows an example of daily vaccine sentiment index from 20210411 to 20210611. In the figure, line "n" is the number of tweets related to vaccine in the random sample in each day.

The vaccine sentiment is estimated by the proportion of positive sentiment in vaccine tweets , that is

$$S_t = \frac{n_{p,t}}{n_t},$$

where S_t is vaccine sentiment on day t, $n_{p,t}$ is the number of tweets with positive sentiment on day t, and n_t is the number of tweets related to vaccine on day t.

The correlation of vaccine sentiment index estimated via three methods is shown in Table 1. The levels of sentiment estimated by sentiment140 and VADER show consistency with correlation of 0.4452. Also, the number of vaccine tweets can also reflect vaccine sentiment to some extent, which can be explained from the aspect of public attention.

5.2 Vaccination Data from CDC

5.2.1 Data Collection. CDC provides thorough data about vaccines in the US by nation and by states (Trends in Number of COVID-19 Vaccinations).

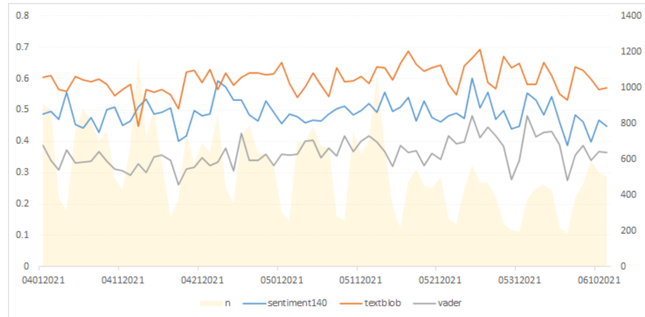

Figure 2: Daily Sentiment Index

Table 1: Correlation of Vaccine Sentiment of Three Methods

correlation	sentiment140	textblob	vader	n
sentiment140	1.0000	0.0366	0.4452	0.1859
textblob	0.0366	1.0000	-0.0811	-0.3489
vader	0.4452	-0.0811	1.0000	0.2425
n	0.1859	-0.3489	0.2425	1.0000

5.2.2 Data Description. Different statistical methods for vaccination situations are available in the daily data file. We choose the following four features as the dataset: "Daily Count People Receiving Dose 1", "7-Day Avg Daily Count Dose 1", "Daily Count of People Fully Vaccinated" and "7-Day Avg Daily Count of People Fully Vaccinated".

5.3 Cases and Deaths Data from CDC

5.3.1 Data Collection. CDC provides data about confirmed cases and deaths in the US by nation and by states (Trends in Number of COVID-19 Cases and Deaths).

5.3.2 Data Description. There are three columns in the case/death data, including "New Cases/Deaths", "7-Day Moving Avg" and "Historic Cases/Deaths", where "Historic Cases/Deaths" refers to the cases/deaths which are excluded from the daily new deaths and 7-day average until they are incorporated into the dataset by the applicable data (explained by CDC).

5.4 Cases Studies Data

5.4.1 Data Collection. To better perform the correlation between the vaccine sentiment/vaccination and the number of cases/deaths, we choose three representative cities: New York, Los Angeles, Houston as case studies. These three cities are with the top 5 population in the United States according to the census led by the Census Bureau on 2019.

We learned that these three cities are composed of 5, 5, and 9 counties, respectively. We can obtain the cities' cases/death and vaccination number by adding the data of the counties. The cases/deaths data by counties is collected from pandemic tracking data system contributed by The New York Times and the county-level vaccination data is obtained from CDC(Vaccinations in the United States, County).

5.4.2 Data Description. Data file of cases/deaths data contains six columns, and there are "date", "county", "state", "fips", "cases" and "deaths", where "fips" refers to a standard geographic identifier, to make an analyst combine this data with other data sets more easily. And files of vaccination data are listed by different classification methods. In this work, we plan to focus on the combination of the number of received vaccines of dose 1 and dose 2.

6 DIFFICULTY

1. There are over 400,000 daily tweets related to COVID-19 in the sample. In this study, we randomly sample 10% tweets to hydrate data more effectively. However, among the hydrated tweets, the amount of vaccine related tweets can not be guaranteed, which may bias the sentiment estimation.

2. For sentiment analysis, it is hard to test the accuracy of classification models on the dataset of vaccine tweets.

3. Both Pfizer and Moderna require two doses to be fully vaccinated. And the vaccine protection efficacy after two injections are not similar. Besides, the vaccination statistics of each community overlap to be impossible to analyze them separately. It may raise the inaccuracy of the impact of vaccine sentiment on cases/deaths.

7 FUTURE PLAN

We have finish the data collection and most work in sentiment analysis. For the final version, there are mainly 3 parts of work.

The first one is to recognize the driver factors of the fluctuation in vaccine sentiment, such as critical events and announcements.

Further, machine learning models will be utilized to build the connection between vaccine sentiment and vaccination rate and cases/death.

Finally, some controlled case studies will be included to show if increasing vaccine sentiment leads to decreases in cases/death in several big cities.

REFERENCES

[1] Sanjida Akter and Muhammad Tareq Aziz. 2016. Sentiment analysis on facebook group using lexicon based approach. In *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*. IEEE, 1–4.

[2] Kashif Ali, Hai Dong, Athman Bouguettaya, Abdelkarim Erradi, and Rachid Hadjidj. 2017. Sentiment analysis as a service: a social media based sentiment analysis framework. In *2017 IEEE International Conference on Web Services (ICWS)*. IEEE, 660–667.

[3] Juan M. Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell. 2021. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration. *Epidemiologia* 2 (8 2021). Issue 3. https://doi.org/10.3390/epidemiologia2030024

[4] Bontempi, Gianluca Ben Taieb, Souhaib Le Borgne, and Yann. 2013. Machine Learning Strategies for Time Series Forecasting. (2013), 62–77. https://doi.org/10.1007/978-3-642-36318-4_3

[5] I Cheng, Johannes Heyl, Nisha Lad, Gabriel Facini, and Zara Grout. 2021. Evaluation of Twitter data for an emerging crisis: an application to the first wave of COVID-19 in the UK. *Scientific Reports* 11, 1 (2021), 1–13.

[6] Wong AHo SOlusanya OAntonini MLyness D. 2021. The use of social media and online communications in times of pandemic COVID-19. , 255-260 pages. Issue 3. https://doi.org/10.1177/1751143720966280

[7] Bijoyan Das and Sarit Chakraborty. 2018. An improved text sentiment classification model using TF-IDF and next word negation. *arXiv preprint arXiv:1806.06407* (2018).

[8] Chedia Dhaoui, Cynthia M Webster, and Lay Peng Tan. 2017. Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing* (2017).

[9] Zulfadzli Drus and Haliyana Khalid. 2019. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science* 161 (2019), 707–714.

[10] Erhu Du, Eddie Chen, Ji Liu, and Chunmiao Zheng. 2021. How do social media and individual behaviors affect epidemic transmission and control? *Science of the Total Environment* 761 (3 2021). https://doi.org/10.1016/j.scitotenv.2020.144114

[11] Facebook. 2021. Facebook Reports Second Quarter 2021 Results. (2021). https://s21.q4cdn.com/399680738/files/doc_news/Facebook-Reports-Second-Quarter-2021-Results-2021.pdf

[12] M.W Gardner and S.R Dorling. 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment* 32, 14 (1998), 2627–2636. https://doi.org/10.1016/S1352-2310(97)00447-0

[13] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* 1, 12 (2009), 2009.

[14] Anees Ul Hassan, Jamil Hussain, Musarrat Hussain, Muhammad Sadiq, and Sungyoung Lee. 2017. Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. In *2017 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 138–140.

[15] Tao Hu, Siqin Wang, Wei Luo, Mengxi Zhang, Xiao Huang, Yingwei Yan, Regina Liu, Kelly Ly, Viraj Kacker, Bing She, and Zhenlong Li. 2021. Revealing public opinion towards COVID-19 vaccines with Twitter Data in the United States: a spatiotemporal perspective. *medRxiv* (2021). https://doi.org/10.1101/2021.06.02.21258233

[16] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

[17] Muhammad Taimoor Khan, Mehr Durrani, Armughan Ali, Irum Inayat, Shehzad Khalid, and Kamran Habib Khan. 2016. Sentiment analysis and the complex natural language. *Complex Adaptive Systems Modeling* 4, 1 (2016), 1–19.

[18] Joanne Chen Lyu, Eileen Le Han, and Garving K Luli. 2021. COVID-19 Vaccine–Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. *Journal of Medical Internet Research* 23 (6 2021). Issue 6. https://doi.org/10.2196/24435

[19] Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

[20] Pooria Taghizadeh Naderi, Ali Asgary, Jude Kong, Jianhong Wu, and Fattaneh Taghiyareh. 2021. COVID-19 Vaccine Hesitancy and Information Diffusion: An Agent-based Modeling Approach. (9 2021).

[21] Fernando P. Polack, Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez, Gonzalo Pérez Marc, Edson D. Moreira, Cristiano Zerbini, Ruth Bailey, Kena A. Swanson, Satrajit Roychoudhury, Kenneth Koury, Ping Li, Warren V. Kalina, David Cooper, Robert W. Frenck, Laura L. Hammitt, Özlem Türeci, Haylene Nell, Axel Schaefer, Serhat Ünal, Dina B. Tresnan, Susan Mather, Philip R. Dormitzer, Uğur Şahin, Kathrin U. Jansen, and William C. Gruber. 2020. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *New England Journal of Medicine* 383, 27 (2020), 2603–2615. https://doi.org/10.1056/NEJMoa2034577 arXiv:https://doi.org/10.1056/NEJMoa2034577 PMID: 33301246.

[22] Naw Safrin Sattar and Shaikh Arifuzzaman. 2021. COVID-19 Vaccination Awareness and Aftermath: Public Sentiment Analysis on Twitter Data and Vaccinated Population Prediction in the USA. *Applied Sciences* 11 (6 2021). Issue 13. https://doi.org/10.3390/app11136128

[23] Jeremy Turiel, Delmiro Fernandez-Reyes, and Tomaso Aste. 2021. Wisdom of crowds detects covid-19 severity ahead of officially available data. *Scientific Reports* 11, 1 (2021), 1–9.

[24] Twitter. 2021. Q1 2021 Letter to Shareholders. (2021). https://s22.q4cdn.com/826641620/files/doc_financials/2021/q1/Q1'21-Shareholder-Letter.pdf

[25] Gregory L. Watson, Di Xiong, Lu Zhang, Joseph A. Zoller, John Shamshoian, Phillip Sundin, Teresa Bufford, Anne W. Rimoin, Marc A. Suchard, and Christina M. Ramirez. 2021. Pandemic velocity: Forecasting COVID-19 in the US with a machine learning and Bayesian time series compartmental model. *PLOS Computational Biology* 17 (03 2021), 1–20. https://doi.org/10.1371/journal.pcbi.1008837

[26] Samira Yousefinaghani, Rozita Dara, Zvonimir Poljak, Theresa M. Bernardo, and Shayan Sharif. 2019. The Assessment of Twitter's Potential for Outbreak Detection: Avian Influenza Case Study. *Scientific Reports* 9 (12 2019). Issue 1. https://doi.org/10.1038/s41598-019-54388-4

[27] E.A. Zanaty. 2012. Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification. *Egyptian Informatics Journal* 13, 3 (2012), 177–183. https://doi.org/10.1016/j.eij.2012.08.002