COVID-19 Vaccination Prediction via Public 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 proposed study, we plan to reveal public opinion towards COVID-19 vaccines with Twitter data and how such sentiment influences vaccination in the US. Our results will provide insight on vaccine campaign and vaccination plan for future epidemics.

KEYWORDS

sentiment analysis, vaccinated case prediction, natural language processing, machine learning

ACM Reference Format:
Jinyang Han, Jingjing Ye, and Qingqing Wu. 2018. COVID-19 Vaccination Prediction via Public Sentiment Analysis on Twitter. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The active users in various social media platforms have skyrocketed in past decades because of the popularity of smartphones and more comprehensive network coverage. Taking the US users as an example, Facebook, Instagram, and Twitter have become the most widely used platforms. Facebook had 1.91 billion daily active users (DAU) on average for June 2020[7], while the DAU of Twitter increased 20% and reached 199 million in the first quarter of 2021[15]. People are increasingly inclined to use social media to record emotions and opinions, providing unprecedented rich resources for studying sentiment propagation and epidemic transmission [6][20].

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 be normal because of all adult vaccination programs released in March 2021. These changes are also reflected in the public social media. To remain connected for work, education, entertainment, and social purposes, people have increased the use of and dependence upon social media platforms[5]. 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[1], combining data of the number of vaccine intake from the CDC to analyze the effect of Twitter posts on public opinion towards vaccines and further predict whether they can affect the vaccination numbers.

Specifically, the problem can be formulated as: Given Twitter users' opinions (in aspects of sentiment, emotions, topics etc.) about vaccines, and the detailed vaccination numbers, build a model to estimate the impact of public sentiment on vaccine intake. Moreover, sentiment difference and vaccination rate can be modeled for comparison across states.

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

2 LITERATURE REVIEW

2.1 Vaccine Sentiment in Twitter

There have been several works analyzing Twitter datasets to reveal vaccine sentiment.

Sentiment analysis, emotion analysis, topic modeling, and other tools are implemented to explore public sentiment and emotions towards COVID-19 vaccines. [9] 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), and announcements of political leaders and authorities (such as Donald Trump tweeting "Great News on Vaccines!") may have potential impacts on public opinion towards COVID-19 vaccines. Further, there exists some geospatial difference in public opinion on vaccines since negative sentiments and emotions are more obvious in some states.

By filtering tweets by keywords associated with different COVID-19 vaccines, for all of the vaccines, people's positive sentiment is $20-25\%$, negative sentiment is around 10% and the rest is neutral. It shows that people still take a positive attitude towards vaccination instead of some adversarial effects of some of the vaccines[14]. [10] has found a consistent result that the sentiment was increasingly positive in general in spite of fluctuations. Also, the most discussed topic related to COVID-19 vaccines may change due to major events about COVID-19 vaccines while opinions about vaccination were the most tweeted topic during the majority of the examination period. As for emotions, trust dominates the discussion continually. The overall percentage of tweets expressing fear is decreasing, indicating that people's fear about pandemics decreased with the progress of vaccine development.

2.2 Vaccine Information Diffusion in Twitter

Additionally, misinformation and information diffusion related to vaccines is also studied. The accounts producing vaccine opposition content were partly Twitter bots or political activists while well-known individuals and organizations generate content in favor of vaccination[19]. The announcement of political leaders and authorities can potentially effect public opinion to vaccines.

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), misinformation and spread of negative messages about COVID-19 vaccines can lead to decline in sentiment and hence cause vaccine hesitancy[9][13].

[12] studies the dissemination of vaccine images on Twitter and distinguish one community against vaccination and the other one community in favor of vaccination in the network. The interaction between the two groups is rare, suggesting that vaccine images are mainly shared within communities. Also members of the anti-vaccine community are more connected than pro-vaccine group and hence they are more likely to retweet and mention each other.

2.3 Social Media and Vaccination Behavior

Many factors can affect people's tendency to vaccination, such as individual characteristics (such as age, gender, income) and vaccine attributes (such as type of vaccines, price of vaccines)[13][18]. Using simulation method, [13] shows that information propagation can change people's minds and their tendency to get vaccinated and suggests considering information diffusion in vaccination simulation.

Besides, [14] uses time series forecasting to predict the percent of people vaccinated in the US.

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.

3 TECHNICAL METHODOLOGY

As shown in Figure 1, our planned work is composed of three ingredients. The first ingredient is vaccine sentiment based on Tweet content analysis. The second ingredient is vaccine behavior based on real vaccination data in the US from CDC. Then machine learning models are implemented to explore the relation between sentiment in Twitter and vaccination.

3.1 Natural Language Processing Algorithm

3.1.1 Sentiment Analysis. Due to lack of pre-labeled dataset, unsupervised lexicon-based approach will be used.

Natural Language Toolkit (NLTK) is the basis for a lot of text analysis done in Python. TextBlob and VADER are useful packages to help make sentiment analysis by dividing sentiment to negative, positive and neutral.

- 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.
- 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. 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.

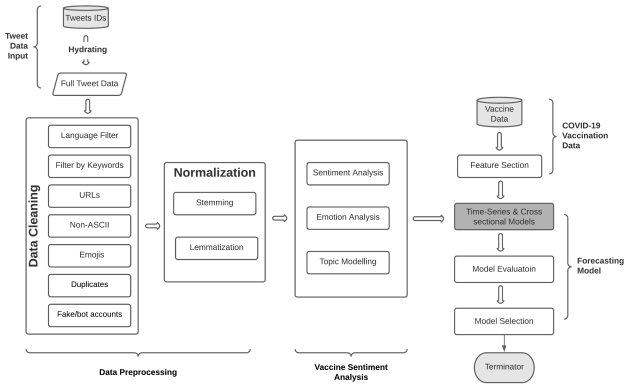

Figure 1: Workflow for Sentiment Analysis and Machine Learning Model

3.1.2 Emotion Analysis. Compared with sentiment analysis to detect positive, negative or neutral opinions, emotion analysis can recognize more types of specific feelings, such as happiness, anger, fear and anticipation. Based on the most comprehensively used dictionary in this area, National Research Council Canada Lexicon[11], emotions about vaccine in Tweets content can be examined.

3.1.3 Topic Modelling. In order to explore the reasons behind the transition of sentiment and emotions, topic modelling can be implemented to classify Twitter content into different themes. One of the most widely used topic models is Latent Dirichlet Allocation (LDA) model[3]. It is developed by [2] and is used in many works to detect topics in social media[9][10].

3.2 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.

3.2.1 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[17].

3.2.2 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[8]. Here we assume MP can get better effect than random forest.

4 EVALUATION

Predication results should always be considered when evaluating whether prediction models perform well. The error metric is a common way to measure the model's quality and provide a method to compare different models. In this research, we want to evaluate our predication using three metrics.

4.1 Mean Squared Error (MSE)

The MSE has ranked top quantitative performance metric for over 50 years [16]. It measures the mean value of the squares of the difference between the predication and true value. We can show this method using Equation (1):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(P_i - T_i)^2, \tag{1}$$

where n is the sample size, T_i is the true value, and P_i is the predication value.

4.2 Root Mean Squared Error (RMSE)

RMSE is a variant of MSE, and it is defined as the square root of MSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - T_i)^2}, \tag{2}$$

where n is the sample size, T_i is the true value, and P_i is the predication value.

4.3 Mean Absolute Percentage Error (MAPE)

MAPE measures the percentage deviation of the forecast value from the actual value, that is

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|P_i - T_i|}{T_i} * 100\%, \tag{3}$$

where n is the sample size, T_i is the true value, and P_i is the predication value.

5 DATA

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

Table 1: Description of CDC vaccine data

No.	Feature	Description
1	Total Doses Administered Daily	Number of COVID-19 vaccine administered per day
2	Daily Count People Receiving Dose 1	Number of people who receive the dose 1 per day
3	Total Doses Administered Cumulative	Cumulative number of COVID-19 vaccine administered
4	People Receiving 1 or More Doses Cumulative	Cumulative number of people receiving more than 1 dose
5	7-Day Avg Daily Count Dose 1	7-day average daily number of people receiving dose 1
6	7-Day Avg Total Doses Daily	7-day average daily number of COVID-19 vaccine per day
7	Total Doses Administered Daily Change	The changing number of COVID-19 administered vaccine per day
8	7-Day Avg Total Doses Administered Daily Change	7-day average daily changing number of COVID-19 vaccine
9	Daily Count of People Fully Vaccinated	Number of people fully vaccinated per day
10	People Fully Vaccinated Cumulative	Cumulative number of people fully vaccinated
11	7-Day Avg Daily Count of People Fully Vaccinated	7-day average daily number of people fully vaccinated

5.1 Vaccination Data from CDC

The Centers for Disease Control and Prevention (CDC) provides thorough data about vaccines in the US by nation and by states (CDC COVID Data Tracker).

Different statistical methods for the total amount of vaccines and increments are available, including: "Total Doses Administered Cumulative", "Total Doses Administered Daily Change","7-Day Avg Total Doses Daily", etc.

Table 1 lists some of the CDC vaccine datasets from which we will choose the most representative attributes to integrate into our model. The time period covered can date back to December 13, 2020.

5.2 Tweets Data from Panacea Lab

As more and more researchers have discovered the ever-increasing number of active users and posts in social media platforms, especially during the epidemic, many laboratories and research institutes, such as Panacea Lab, scrape tweets that contain keywords including "COVID-19" or "vaccine". Besides, they also provide the free download link to share the data set.

Data provided by the Panacea Lab is one of the most widely used Covid-19 Twitter chatter datasets in research (Panacea Lab). The dataset includes tweets data from March 22, 2020 until now.

The database we downloaded was collected up to October 2, 2021, for a total of 561 days. The file contains three columns, "tweet_id",

"date" and "time", which represent the tweet id of the collected post, posting date and posting time respectively. Additionally, we need to use "Hydrator", a open-source hydration software, to recover the full tweet content and geolocation data based on the tweet IDs.

Specifically, tweet content, posting time as well as locations are most relevant to our research since we plan to extract sentiment information by state using content analysis.

6 EXPECTATION

At the end of the project, we want to accomplish two main goals.

Firstly, we will construct a COVID-19 vaccine sentiment monitoring tool based on Twitter. The input of the model is Tweets data regarding to vaccine. After sentiment and emotion analysis, the model can generate a COVID-19 vaccine sentiment map, which shows public attitude towards COVID-19 vaccine and the level of sentiment by state. Also public attention in vaccine and the potential drivers of fluctuation in sentiment can be investigated based on topic modelling. This model can be effective in detecting the diffusion of panic and misinformation related to COVID-19 vaccine in Twitter, with which measures can be taken to tackle possible negative influence.

Secondly, a vaccine tendency model can be built based on public attention to COVID-19 vaccine. The model can be used to detect factors that may influence people's willingness to vaccinate against COVID-19 in different states. In turn, the conclusions will help to formulate solutions to improve vaccine rates.

Overall, our results will give hints on strategies to improve coverage of vaccinated population to realize community immunity.

7 DIVISION OF LABORS

After discussion, each of our members has the following expected task division. Jinyang Han and Qinqing Wu will focus on natural language processing of raw data and evaluate the NLP performance. Jingjing Ye will be responsible for prediction of vaccination using machine learning and evaluation of model results.

The overall expected timeline of our project is shown in Figure 2.

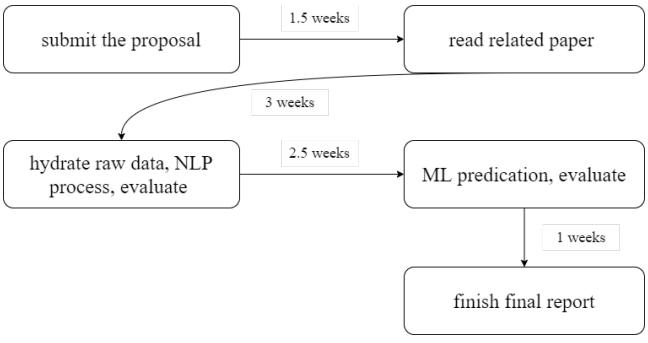

Figure 2: Expected Timeline of Activities

REFERENCES

[1] 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

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.

[3] Avrim Blum, John Hopcroft, and Ravindran Kannan. 2016. Foundations of data science. *Vorabversion eines Lehrbuchs* 5 (2016), 5.

[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] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] Mohammad Saif M and Turney Peter D. 2013. Nrc emotion lexicon. *National Research Council, Canada* 2 (2013).

[12] Elena Milani, Emma Weitkamp, and Peter Webb. 2020. The visual vaccine debate on Twitter: A social network analysis. *Media and Communication* 8, 2 (2020), 364–375.

[13] 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).

[14] 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

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

[16] Zhou Wang and Alan C. Bovik. 2009. Mean squared error: Lot it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine* 26 (2009), 98–117. Issue 1. https://doi.org/10.1109/MSP.2008.930649

[17] 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

[18] Fulian Yin, Zhaoliang Wu, Xinyu Xia, Meiqi Ji, Yanyan Wang, and Zhiwen Hu. 2021. Unfolding the Determinants of COVID-19 Vaccine Acceptance in China. *Journal of Medical Internet Research* 23 (1 2021). Issue 1. https://doi.org/10.2196/26089

[19] Samira Yousefinaghani, Rozita Dara, Samira Mubareka, Andrew Papadopoulos, and Shayan Sharif. 2021. An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *International Journal of Infectious Diseases* 108 (7 2021). https://doi.org/10.1016/j.ijid.2021.05.059

[20] 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