Public Attitudes towards the COVID-19 Vaccine in the U.S.A. based on Tweets

Hangqian Li
École polytechnique fédérale de Lausanne
Lausanne, Switzerland
hangqian.li@epfl.ch

Junhong Li
École polytechnique fédérale de Lausanne
Lausanne, Switzerland
junhong.li@epfl.ch

Zijun Cui
École polytechnique fédérale de Lausanne
Lausanne, Switzerland
zijun.cui@epfl.ch

Abstract

To explore the attitudes of different geographic regions of the United States towards COVID-19 vaccines, and the attitudes towards COVID-19 vaccines from different companies, we use a week of US Tweet to analyze the sentiment positive level in the two dimensions and explore the reasons behind the phenomenon. We have some interesting findings, such as Pfizer has lower positive level in that week compared with AstraZeneca.

CCS Concepts: • **Social and professional topics** → *Computing / technology policy*; *Geographic characteristics*.

Keywords: Twitter, COVID-19 vaccine, public attitude, U.S.A.

ACM Reference Format:
Hangqian Li, Junhong Li, and Zijun Cui. 2021. Public Attitudes towards the COVID-19 Vaccine in the U.S.A. based on Tweets. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2021, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

1 Introduction

The impact of the COVID-19 health crisis has marked it as an unique milestone in the history of disease outbreaks [2]. The development and production of vaccines are urgent while time is limited. Although a variety of COVID-19 vaccines have been put into use in a wide range of countries, global distrust of vaccines has been around since last year, which is one of the unfavorable factors for controlling the pandemic [7]. This kind of hesitancy or worries about vaccination can be summed up in one term - Vaccine hesitancy.

There have been many public opinion surveys about the willingness to vaccinate the COVID-19 vaccines all over the world. Malik Sallam's research synthesized the results of surveys on vaccine hesitancy from different countries around the world [9]. It shows that the vaccine acceptance rate of the United States is at a low level globally until the end of last year, which was only 56.9%. An official statistics on vaccine acceptance rates in the U.S. from April 14 to April 26 this year is shown in Figure 1 [1]. It can be seen that in the east coast cities of the U.S., such as Boston, New York, and the west coast cities like Los Angeles, San Francisco, the acceptance rate is relatively high. The regions with the relatively low acceptance rate appear in the central region, as well as parts of the eastern region.

It is worth noting that most previous surveys on vaccine hesitancy relied on questionnaires. This method is valid but also has some shortages such as the samples collected are limited. In comparison, social media platforms can provide a larger range of up-to-date samples that could be used for analysis. Thus, social media data like Tweets can be helpful to know the latest public attitudes. In addition, the fact that the health tolls of COVID-19 in the U.S. have been among the highest in the world[4], while the vaccine acceptance rate has been low makes American Tweets related to COVID-19 vaccines worthy of analysis.

The goal of this research contains two aspects: one is to know the latest attitudes of the American people in different regions towards the COVID-19 vaccine; another is to learn the latest vaccine hesitancy in the U.S. for COVID-19 vaccines from different companies. For this purpose, six vaccines that are currently mainstream in the world were selected for analysis.

2 Methodology

The main procedure of this research is presented as following:

1. Firstly, we collect a labelled dataset in which the text is labeled as positive or negative and a large number of Tweets.
2. Then we preprocess the labelled data and the Tweets in the same procedure we mentioned below in 3.2 to get a similar structure.
3. In order to train and test the model, we extract features from both labelled data text and the Tweets text to

Woodstock '18, June 03–05, 2021, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

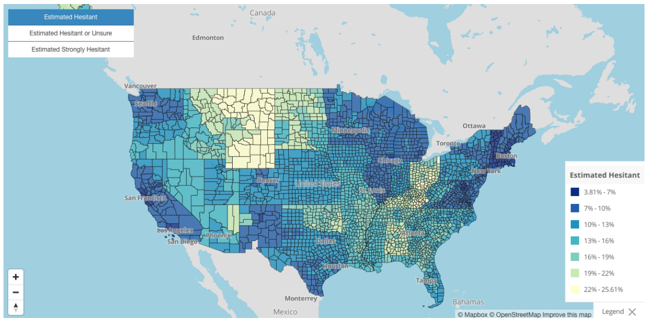

Figure 1. The above map shows estimates of COVID-19 vaccine hesitancy rates using data from the U.S. Census Bureau's Household Pulse Survey (HPS) from April 14, 2021 to April 26, 2021 [1].

vectorize them, which then can be used as the input of the model.

4. Finally we train our model with the training dataset. Two widely used models are tested in this stage: the Logistics Regression model and the Naive Bayes model.

2.1 Data collection

2.1.1 Tweets data collection. We collect two sets of Tweets in this research with Tweepy library. One is based on different locations and used to understand the latest attitudes of the American people in different regions towards the COVID-19 vaccine. The other is based on different vaccines and used to learn the latest vaccine hesitancy in the U.S. for COVID-19 vaccines of different companies (Astrazeneca, Pfizer, Janssen, Sputnik, Moderna, CoronaVac).

2.1.2 Labelled data collection. We use a csv dataset in which 1.6 million Tweets are already coded into two categories by hand [10]. Two variables are used, which are the Tweet text and the sentiment label target. The sentiment label is comprised of the integer values 0 as negative and 4 as positive. We split the labelled dataset into 80% training set and 20% test set.

2.2 Data prepossessing

The main purpose of the prepossess is to have a standard represent for the text and to remove some meaningless symbols and words. 9 steps are shown as following:

1. Filter out non-English Tweets
2. Set the Tweets to lowercase
3. Filter URL links
4. Remove username, @, and # from Tweets
5. Filter emoticons
6. Filter punctuation
7. Lemmatization: convert a word to its base form
8. Remove stopwords
9. Remove short words (length strictly less than 3 characters)

Firstly, we filter out the non-English Tweets and set all the remaining Tweets to lower case.
Secondly we filter out meaningless text and symbols like URL links, usernames, punctuation, emoticons and some other symbols.
Then we use the lemmatization method to convert the words to their base forms like cars to car to have a standard represent for the text.
Finally, to reduce the amount of words and features, we remove stop words and short words whose length is strictly less than 3 characters.

2.3 Feature extraction

In order to vectorize the input text, we firstly extract features from the original text. An obvious difference that can be made among terms is with respect to the frequency of occurrence in a document. Thus a weighting scheme for documents can be defined by considering the relative frequency of terms within a document. The term frequency is normalized with respect to the maximal frequency of all terms occurring within the document. But we have to take into account not only the frequency of a term within a document when determining the importance of the term for characterizing the document, but also the discriminative power of the term with respect to the document collection as a whole. For this purpose, the inverse document frequency is computed and included into the term weight. Hence, we choose term frequency–inverse document frequency index (TF-IDF).

$$tf(i,j) = \frac{freq(i,j)}{max_{k \in T}(freq(k,j))} \tag{1}$$

$$idf(i) = log(\frac{n}{n_i}) \tag{2}$$

$$w_{i,j} = tf(i,j) \times idf(i) \tag{3}$$

freq(i,j) means number of term k_i occurring in document d_j. n_i means number of documents in which term k_i occurs in the equation and n is total number of documents.

2.4 Classifier

To label the sentiment of Tweets is a binary classification task, thus we choose two widely used models: the Logistics Regression model and Naive Bayes model.

Logistic regression permits one to type a multivariate regression relationship between a dependent variable and several independent variables. Logistic regression, which is one of the multivariate analysis models, is helpful for forecasting the presence or absence of a characteristic or outcome based on the values of a set of predictor variables. The advantage of logistic regression is that, through the addition of a suitable link function to the usual linear regression model, the variables may be either continuous or discrete, or any combination of both types and they do not necessarily have normal

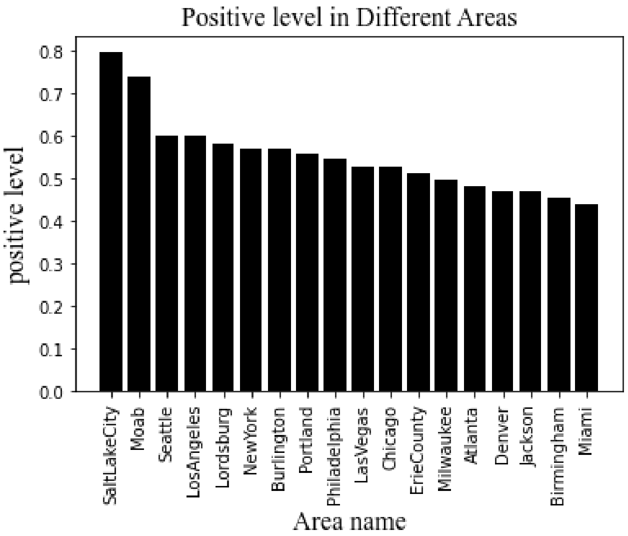

Figure 2. Twitter public attitude level in different areas, sorted in descending order.

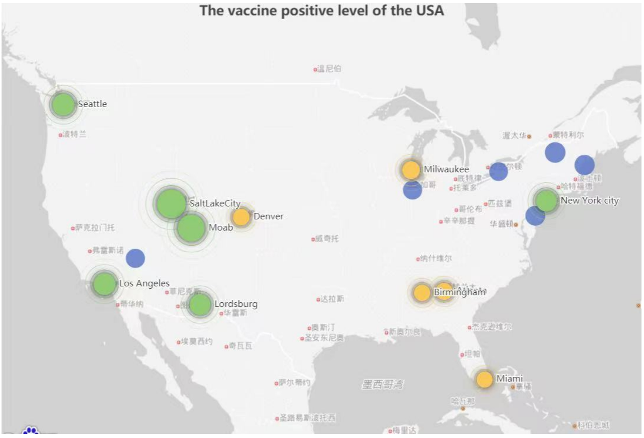

Figure 3. Twitter public attitude level in different areas, marked on the U.S.A. map. Green points mean the highest-5 attitude positive level cities, yellow points means the lowest-5 attitude positive level cities.

distributions [11]. We used Logistics Regression function in sklearn library and get a 77% accuracy for the model.

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. The classifier assigns the mostly likely class to given a example described by its feature vector [6]. We used multinomiaNB function from sklearn library and get a 76% accuracy. There is no significant difference between the performance of two models. We finally choose logistics model with 77% accuracy to classify our Tweets text.

3 Results and Discussion

3.1 Attitude positive level in different geographical regions

We first calculate the Twitter public attitude level in different regions of the U.S.A, which could be seen in Figure 2. The highest attitude positive level is in Salt Lake City, which is 79.6%. For some big cities, such as Seattle, Los Angeles and New York, they have over 57% attitude positive level. At the same time, some small towns, such as Moab, Lordsburg, also have high attitude positive level, ranking 2nd and 5th among the 18 cities. While the lowest level is in Miami, which attitude positive level is only 44.0%. The rest four lowest regions are Biemingham, Jackson, denver, and Atlanta, which are 45.4%, 46.9%, 46.9%, 47.9% respectively.

In order to find the public vaccine attitude characteristics of different geographic regions of the United States, we draw a map of the United States and marked the 5 most positive regions and the 5 most negative regions for the COVID-19 vaccine (see Figure 3). Green points mean the highest-5 attitude positive level cities, yellow points means the lowest-5 attitude positive level cities. From the figure, it seems that people in the West and East has higher attitude positive level, but in the central east region of the U.S.A., the positive level is relatively low. It is almost the same as the previous vaccine hesitancy heat map, whose data is collected by questionnaires (see Figure 1).

3.2 Correlation between public attitude and vaccinated rate, COVID-19 positive rate.

From Covid Act Now, which is an independent nonprofit founded by volunteers to help people make informed decisions by providing timely and accurate data about COVID in the U.S., we can get the vaccinated and COVID-19 positive rate of different regions (see Table 1) [5]. The highest-five COVID-19 positive rate cities have a mean attitude positive level of 0.62; while the lowest-five COVID-19 positive rate cities have a mean attitude positive level of 0.56. As for the vaccinated rate, the highest-five cities have a mean attitude positive level of 0.55; while the lowest-five have a mean attitude positive level of 0.50. It seems that the attitude positive level is positively correlated with both COVID-19 positive vaccinated rate and COVID-19 positive rate.

To verify if the above conjecture is correct, we do hypothesis testing by calculating the Pearson correlation coefficients. The Pearson correlation coefficient measures a linear relation and can be highly sensitive to outliers [8]. The Correlation between public attitude positive level towards vaccine and COVID-19 vaccinated rate: There is a small (0.38), and not significant ($p = 0.12 > 0.05$) positive correlation (see Figure 4). The Correlation between public attitude positive level towards vaccine and COVID-19 positive rate: There is a small

Table 1. Attitude positive level, COVID-19 positive rate, and vaccinated rate in different regions

Regoin	Attitude positive level	COVID-19 positive rate	Vaccinated rate
Salt Lake City	0.796053	0.034	0.494
Moab	0.740000	0.096	0.526
Seattle	0.601643	0.026	0.548
Los Angeles	0.600000	0.003	0.552
Lordsburg	0.580645	0.011	0.412
New York	0.570470	0.005	0.645
Burlington	0.569606	0.016	0.553
Portland	0.557369	0.011	0.654
Philadelphia	0.546928	0.025	0.425
Las Vegas	0.528662	0.037	0.419
Chicago	0.525735	0.019	0.581
Erie County	0.511798	0.016	0.563
Milwaukee	0.495055	0.016	0.460
Atlanta	0.479473	0.023	0.364
Denver	0.469394	0.015	0.447
Jackson	0.468707	0.021	0.431
Birmingham	0.453774	0.055	0.373
Miami	0.440252	0.022	0.378

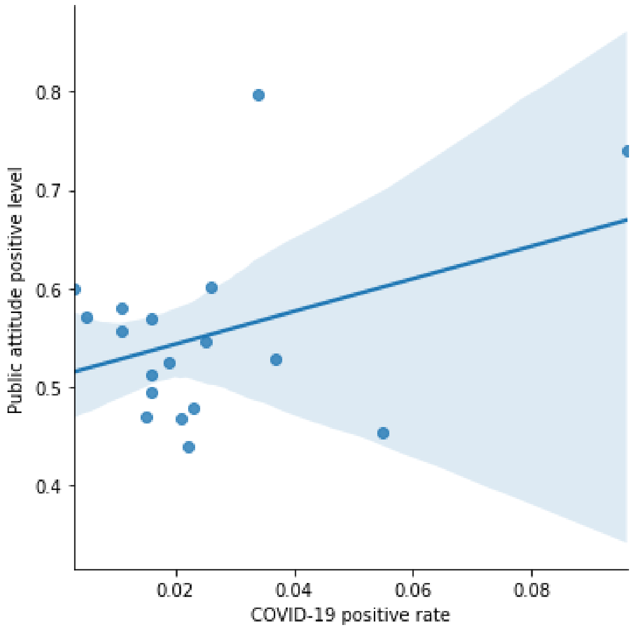

Figure 4. The correlation between Twitter public attitude level and COVID-19 vaccinated rate.

(0.39), and not significant (p = 0.11 > 0.05) positive correlation (see Figure 5). The p-value is not low, so the correlation is not statistically significant.

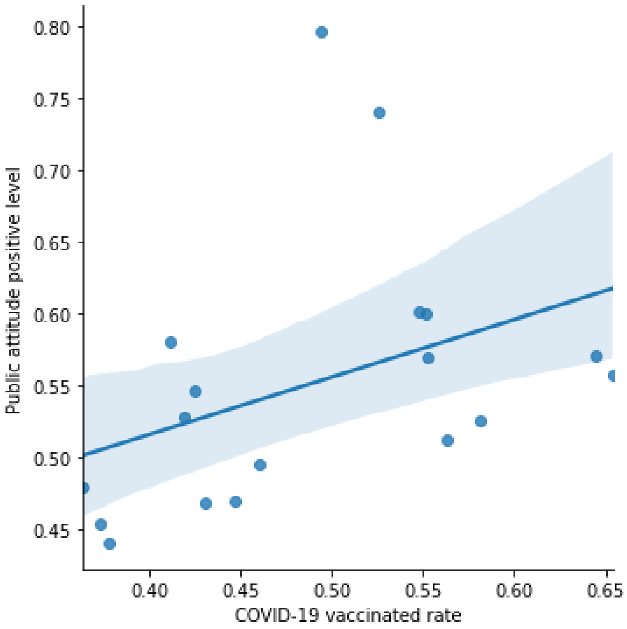

Figure 5. The correlation between Twitter public attitude level and COVID-19 positive rate.

3.3 Twitter attitude positive level towards different vaccines in the U.S.A.

We also calculate the Twitter public attitude level towards different vaccines in the U.S.A., which could be seen in Figure 6. AstraZeneca ranks highest among those six vaccines, with around 0.673 attitude positive level; Moderna ranks second

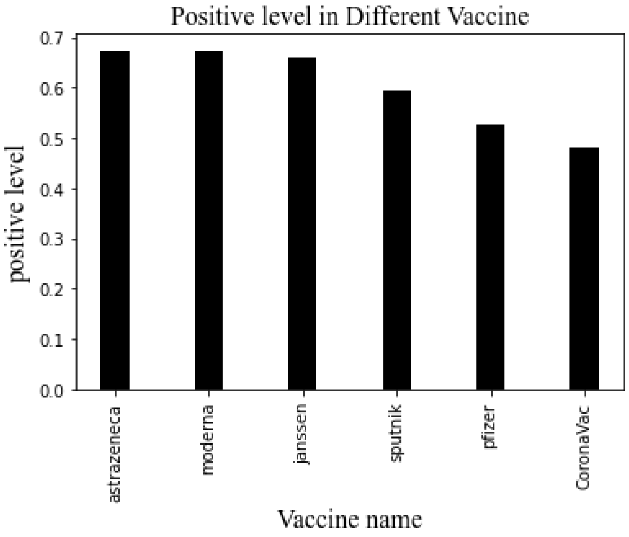

Figure 6. Twitter public attitude level towards different vaccines.

highest among those six vaccines, with around 0.671 attitude positive level; Janssen, and Sputnik rank third and fourth with 0.660 and 0.595 separately; Pflizer ranks fifth with 0.528 attitude positive level, which is surprising to us; CoronaVac ranks lowest with 0.481 attitude positive level.

From Yale Medicine, we know that Pfizer-BioNTech has a better efficacy rate, which is 95%, in preventing COVID-19 than other vaccines. While Oxford-AstraZeneca might some serious side effects, such as blood clot [3]. But in the previous result of sentiment analysis shows some interesting phenomena such as Pfizer-BioNTech related Tweets are more negative than the Oxford-AstraZeneca related Tweets. Next, we will explore the potential reasons behind these interesting phenomena.

3.3.1 Pfizer-BioNTech.

We check some of Pfizer's negative comments and generate a corresponding negative Tweet word cloud (see Figure 7).

And here are some negative remarks about Pfizer-BioNTech:

- I've got nothing nice to say either. Annastacia Palaszczuk is over 50, just like me, but I only qualify for the AstraZeneca vaccine, which carries more risk. Don't tell this is blatant discrimination because it is.
- My reluctance to be vaccinated is simple...lack of choice...at 54 I would like the choice to receive the Pfizervaccine...why…
- Has anyone experience leg pain following the 2nd dose of Pfizer? My 13 year old is having increasing pain in his left thigh. Seems weird.
- Surprise the virus, mutate first!
- Bloody hell! Most of the people I've spoken with have had the Pfizer jab

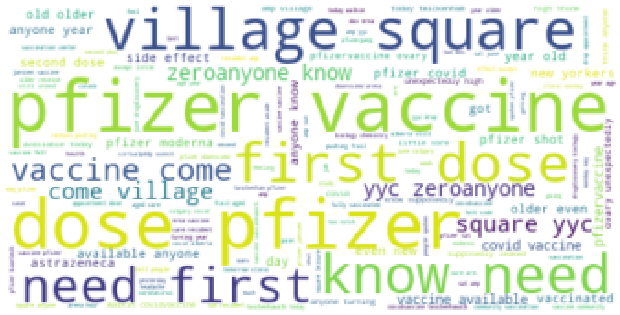

Figure 7. Word cloud of negative Tweets of **Pfizer-BioNTech** in the U.S.A.

It can be seen that many negative comments are not about Pfizer's effect, but some negative events surrounding Pfizer, such as some unfair vaccine distribution rules, discovering that friends around the user have been vaccinated but he/she have not been vaccinated. And the reaction after being prevented by the unfair vaccine distribution rules to vaccinate Pfizer, on the other hand, could reflect the positive affirmation of Twitter users to Pfizer.

3.3.2 Oxford-AstraZeneca.

Oxford-AstraZeneca ranks the highest among the six vaccines with around 0.673 attitude positive level, which is not the same as our expectation. The word cloud of Oxford-AstraZeneca's positive comments could be seen in Figure 8.

Some positive remarks about Oxford-AstraZeneca:

- Thank you Denmark for pledging the direct donation of these AstraZeneca vaccines! Nepal will be forever grateful for this kind gesture.
- 'Health Canada just extended the expiration of astrazenecavaccine by a month. Sensible.
- 'Just got my 2nd AstraZeneca shot tonight. But this is good info for those in the hunt!!!!
- Sure glad I got my second doze of astrazenecavaccine yesterday!
- Great joy! Today at 12:24pm I had my 2nd AstraZeneca Covid vaccination jab! I was due to have it on 12th June but yesterday I was summoned to have an earlier appointment.

By analyzing the actual positive comments, we found that the location of some Tweets might be Canada. When we collect the data in practice, we draw a circle with a radius of 1500 miles in the center of the United States and filter out Tweets in Spanish. There may still be some Tweets from Canada and Mexico, which cannot be filtered. So although the AstraZeneca vaccine cannot be vaccinated in the United States, it can be vaccinated in neighboring countries.

We also found that in positive comments of AstraZeneca, there are many news about sharing the AstraZeneca vaccine with other countries, such as White House lays out plan to

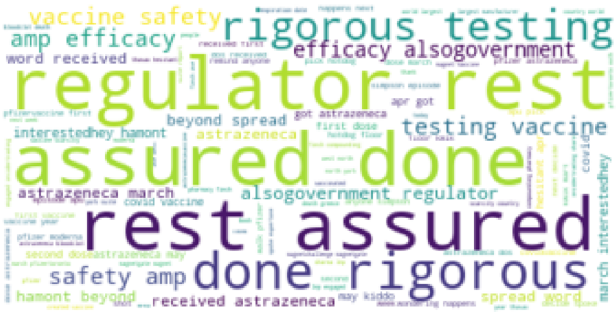

Figure 8. Word cloud of negative Tweets of **Oxford-AstraZeneca** in the U.S.A.

share millions of Covid doses with poorer nations, Japan to donate AstraZeneca vaccine doses to Philippines. There are a bunch of thank you Tweets about AZ's donation in the data sample. Meanwhile some people who have been vaccinated with AstraZeneca will post some positive Tweets. These could explain why AstraZeneca related Tweets are relatively positive in some sense.

3.3.3 China's Sinovac.

China's Sinovac ranks the lowest in the U.S.A, which is not very surprising to us, because of the ideological and trade conflicts between the United States and China. To look into details what the negative comments are, we generate a China Sinovac negative Tweet word cloud (see Figure 9).

Some negative remarks about Sinovac:

- Any Doctor? Yesterday my cousin got sinovac shot now today he have some allergic rashes on his face and no other reaction…'
- Are you getting all the info on COVID19? If you read Reuters, don't miss out on what 15 other sources have to say Sinovac Coronavirus Vaccines WorldHealthOrganization
- WHO Requests that Sinovac Share more Vaccine Data to Consider EUA.

We can see that some of the negative Tweets about Sinovac are about some minor side effects, and some are news about WHO approving Sinovac COVID shot. It is also worth mentioning that the Sinovac's whole data sample is only 135 Tweets, much smaller comparing to pfizer 3371 and moderna 1307, which might cause some bias to the public attitude analysis.

3.3.4 Moderna.

Moderna ranks second among the six vaccines with around 0.671 attitude positive level. The word cloud of Moderna's positive comments could be seen in Figure 10.

Some positive remarks about Moderna:

- Moderna has recently announced their COVID19 vaccine is safe and effective for patients aged 12-17. So relieved Mom's gotten her second shot. Doing much better with Moderna this round. We came ready with Tylenol, Gravol, Coffee and some snacks. Just a bit of soreness.
- Pfizer and Moderna to launch new mRNA product to vaccinate against lightning strikes.
- The CDCgov recently confirmed the safety of the Pfizer and Moderna vaccines for expectant mothers and babies.

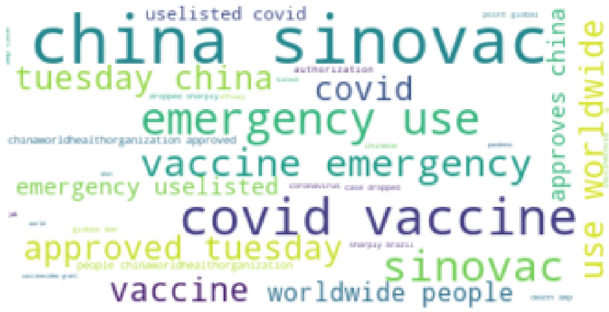

Figure 9. Word cloud of negative Tweets of **China Sinovac** in the U.S.A.

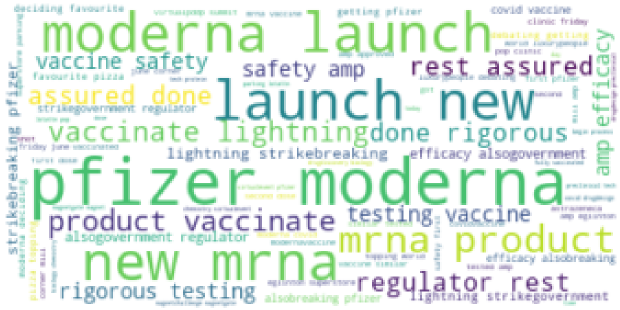

Figure 10. Word cloud of negative Tweets of **Moderna** in the U.S.A.

It can be seen that the positive Tweets of Moderna is basically the positive comments after vaccination, or the affirmation of Moderna vaccine by some organizations.

4 Conclusion

The public hesitancy towards to COVID-19 vaccine could come from various internal and external reasons, and sometimes the attitudes are very subjective and vary from person to person. It requires government policy makers and media workers to adopt more appropriate measures to build COVID-19 vaccination trust among the general public through public messages and take more effective methods when implementing vaccine policies.

There are still many limitations in our work that need future work to improve. Due to the limitation of Tweet API and computing power, we only selected a short period (2021-May-29 to 2021-June-6) of time but did not choose a longer period time span. We only sampled 18 regions in the U.S.A., rather than the whole U.S.A., and there are several under-samplings in the central area. As for some regions and vaccines, there are only a few hundred Tweets. Many of these Tweets might be retweets. A large amount of similar Tweet might affect the accuracy of the results. Some remarks are not about the attitude towards the vaccines themselves, might be affected by some specific events about the COVID-19 vaccine. In future work, more data needs to be collected, and more sentiment analysis models need to be applied to more accurately show the US attitude towards the new crown vaccine.

References

[1] Centers for Disease COntrol and Prevention. 2021. estimates of vaccine hesitancy for COVID-19. https://data.cdc.gov/stories/s/Vaccine-Hesitancy-for-COVID-19/cnd2-a6zw/. Accessed: 2021-06-14.

[2] M. Haghani and M. Bliemer. 2020. Covid-19 pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across SARS, MERS and 2019-nCoV literature. *Advance online publication* (2020), 1–32. https://doi.org/10.1007/s11192-020-03706-z

[3] KATHY KATELLA. 2021. Comparing the COVID-19 Vaccines: How Are They Different? https://www.yalemedicine.org/news/covid-19-vaccine-comparison. Accessed: 2021-06-18.

[4] Jagdish Khubchandani, Sushil Sharma, James H. Price, Michael J. Wiblishauser, Manoj Sharma, and Fern J. Webb. 2021. COVID-19 Vaccination Hesitancy in the United States: A Rapid National Assessment. *Journal of Community Health* 46, 2 (Apr 2021), 270–277. https://doi.org/10.1007/s10900-020-00958-x

[5] Covid Act Now. 2021. *U.S. COVID Risk & Vaccine Tracker.* Retrieved July 8, 2021 from https://covidactnow.org/?s=1933851

[6] Irina Rish et al. 2001. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, Vol. 3. 41–46.

[7] Michelle Roberts. 2019. Vaccines: Low trust in vaccination 'a global crisis'. https://www.bbc.com/news/health-48512923. Accessed: 2021-06-18.

[8] Ronald Rousseau, Leo Egghe, and Raf Guns. 2018. Chapter 4 - Statistics. In *Becoming Metric-Wise*, Ronald Rousseau, Leo Egghe, and Raf Guns (Eds.). Chandos Publishing, 67–97. https://doi.org/10.1016/B978-0-08-102474-4.00004-2

[9] Malik Sallam. 2021. COVID-19 Vaccine Hesitancy Worldwide: A Concise Systematic Review of Vaccine Acceptance Rates. *Vaccines* 9, 2 (Feb 2021), 160. https://doi.org/10.3390/vaccines9020160

[10] Gaurav Singhal. 2021. *training.csv.* Retrieved July 8, 2021 from https://www.dropbox.com/s/du1z2m910a68ehk/training.csv?dl=0

[11] A Yalcin, Selçuk Reis, AC Aydinoglu, and T Yomralioglu. 2011. A GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods for landslide susceptibility mapping in Trabzon, NE Turkey. *Catena* 85, 3 (2011), 274–287.