

# Social Media Analytics Project 2021

How US Government could  
manage Covid-19 Vaccination  
through Tweets

TEACHER: Nuno António

STUDENTS:

André Calado M20201263

Cátia Simões M20201274

Jorge Alves M20201192

Joana Fonseca M20201228

Marius Hessenthaler E20201824



---

## Table of Contents

<b><i>Introduction</i></b> .....	<b>3</b>
<b><i>Data Extraction</i></b> .....	<b>4</b>
<b><i>Data Preprocessing</i></b> .....	<b>5</b>
General .....	5
Frequency Analysis Preprocessing .....	5
Keyword Extraction Preprocessing .....	5
Sentiment Analysis Preprocessing .....	5
Hate Speech Analysis Preprocessing .....	6
Topic Modeling Preprocessing.....	6
Named Entity Recognition Preprocessing.....	6
<b><i>Individual Analysis</i></b> .....	<b>7</b>
Frequency Analysis .....	7
Keywords Extraction .....	7
Sentiment Analysis .....	7
Hate Speech Analysis .....	7
Topic Modeling.....	8
Named Entity Recognition.....	9
<b><i>Combined Analysis</i></b> .....	<b>11</b>
<b><i>Conclusions</i></b> .....	<b>12</b>
<b><i>Suggestions</i></b> .....	<b>15</b>
<b><i>Bibliography</i></b> .....	<b>16</b>
<b><i>Annexes</i></b> .....	<b>16</b>

---

# Introduction

More than one year after the pandemic started the end of the pandemic in the United States (US) is within reach with various COVID-19 vaccines and rapidly increasing availability in the western world. After desperately waiting and strong government efforts for developing and producing vaccines in 2020 the vaccination campaign starts to stagnate after not even half of the US population is fully vaccinated. It is in the governments and everyone's interest to boost vaccination rates for individual protection as well as herd immunity that is estimated at 85% of fully vaccinated population (Robert-Koch-Institut, 2021) to protect the weakest of the society.

To come closer to this goal the government needs to understand the people's concerns about the vaccination and manage their trust in vaccines and satisfaction with the government's response to for acceptance of upcoming measures. As a marketing analytics firm that was instructed by the United States of America government to analyze the social media perception of their SarS-CoV-2 vaccination campaign we will use Text and Social Media Analytics methods to do so. The goal is to find reactions, problems, and criticism regarding the execution of the vaccination campaign but also regarding the vaccines itself.

We decided to use Twitter as the Social Media platform since posts are easy to extract and users address more political topics than on others. Further we decided to use Frequency Analysis and Keyword Extraction methods to identify the most relevant n-grams and key sentences that are mentioned in the extracted data. For a combined statistical analysis of the corpus regarding certain topics, actors and sentiments, we apply multiple recognition, clustering and classification methods. We apply Sentiment Analysis for an indicator of a Tweets opinion. We deemed hate speech analysis as important as Twitter posts are often subject to hateful posts that can be another opinion indicator or also used for further filtering Tweets. We use Named Entity recognition to identify and count relevant entities. Topic Modeling is used to cluster the corpus into a fixed number of topics that can be used subsequently to gain more insights into the topic itself.

---

# Data Extraction

To extract relevant Post from Twitter (Tweets) we used the Twitter API via the python package Tweepy. The extraction was based on the lecture's API Demo Notebook. The most important alteration giving the Twitter search query. It had to satisfy the requirements that the Tweet addresses a government official or organization which we realized by filtering for mentions of at least one of a selected number of Twitter accounts. Second it needed to be about the topic of vaccination which we filtered by containing at least one of a number of vaccination (brand) search terms. Additionally we wanted opinions of basic citizens instead of politician or media, therefore we excluded Tweets by verified users. This resulted in the following search query.

```
'(@joebiden OR @USAGov OR @POTUS OR @HHSvaccines OR @US_FDA OR @CDCgov OR @WhiteHouse) (vaccine OR vaccination OR vax OR moderna OR AstraZenca OR Biontech OR JNJ) -(is:verified)'
```

We extracted a total of 22.543 Tweets in the time period of the 16<sup>th</sup> May until 10<sup>th</sup> June 2021.

# Data Preprocessing

## General

Before continuing with the different Data Analysis steps and specific preprocessing we applied some general preprocessing on the data. Consisting of the following:

1. Removing Hashtags and Mentions and put them into separate columns
2. Removing Urls
3. Text replacement (such as white spaces and vax with vaccine)
4. Removing tweets from the user threadreaderapp (a bot to summarize threads of Tweets)
5. Removing HTML Tags
6. Removing line breaks
7. Removing consecutive spaces
8. Removing special characters (with ' [^\x00-\xfd] ')

Note: Steps 5. – 8. are in the individual analysis notebooks but applied in each of them.

Regarding step 1 we first proceeded with only removing the “#” itself but not its text. However due to the nature of hashtags that often multiple words are connected without spaces the following algorithms delivered undesired results. As a result, we removed the hashtags completely.

## Frequency Analysis Preprocessing

For the frequency analysis the following preprocessing steps are performed:

1. Removing emoticons and emojis (with the emot and emoji packages)
2. Removing special characters (with '\?|\.|\\!|\\;|\\.|\\\"|\\,|\\(|\\)|\\&|\\:|\\-|\\')
3. Removing numbers
4. Converting to lower case
5. Stop word removal
6. Stemming

## Keyword Extraction Preprocessing

For the keyword extraction the following preprocessing steps are performed:

1. Removing emoticons and emojis (with the emot and emoji packages)
2. Converting to lower case

## Sentiment Analysis Preprocessing

For the sentiment analysis only numbers are removed because most other features (like upper case or emojis) can influence the sentiment.

---

## Hate Speech Analysis Preprocessing

For the hate speech analysis the following preprocessing steps are performed:

1. Removing emoticons and emojis (with the emot and emoji packages)
2. Removing special characters (with '\?|\.|\\!|\\;|\\.|\\\"|\\,|\\(|\\)|\\&|\\:|\\-|\\')
3. Removing numbers
4. Word tokenization

## Topic Modeling Preprocessing

For the topic modeling the following preprocessing steps are performed:

1. Removing emoticons and emojis (with the emot and emoji packages)
2. Removing special characters (with '\?|\.|\\!|\\;|\\.|\\\"|\\,|\\(|\\)|\\&|\\:|\\-|\\')
3. Removing numbers
4. Converting to lower case
5. Word tokenization
6. Stop word removal
7. Lemmatization

## Named Entity Recognition Preprocessing

For the named entity recognition only emoticons and emojis are removed with the emot and emoji packages because other features can help to identify entities in the text.

---

# Individual Analysis

## Frequency Analysis

The frequency analysis is based on the lecture's notebook and finds the most frequently used uni-, bi- and trigrams in the Tweets.

## Keywords Extraction

For the keyword extraction analysis we use the Rapid Automatic Keyword Extraction (RAKE) method based on the lecture's notebook. For the Tweet data set RAKE identifies many unrelated, spam or foreign language keywords. We suspect this is the case because those or similar Tweets were posted (by copy-and-paste) multiple times to draw attention. We therefore manually scan the list of keywords to identify relevant keywords. The complete list can be found in the project's files and the manually selected sentences are presented in the Conclusion section.

## Sentiment Analysis

The sentiment analysis is based on the lecture's notebook which uses the Vader Sentiment Intensity Analyzer. We bin the Tweets based on the sentiment compound score into the classes:

- positive: score  $\geq 0.05$
- neutral:  $-0.05 < \text{score} < 0.05$
- negative: score  $\leq -0.05$

## Hate Speech Analysis

For the hate speech analysis, we first decided to build a classifier with the state-of-the-art BERT pretrained language model. BERT performs well on all kind of natural language processing tasks and should because of its context-sensitive representation especially well for complex task as hate-speech detection (Devlin, 2019). For the training we use an annotated general Twitter hate speech data set which is available on Kaggle<sup>1</sup>. The dataset is already partially preprocessed, but we additionally remove all special characters and emojis and finally tokenized with the BERT tokenizer with the BERT uncased base model. We randomly split the data set into 80 % training, 10 % validation and 10 % test set. Finally, we train the model on the training set which achieves an f1-score of 0.76 with contains hate speech as the positive class and an overall accuracy of 0.97. These results would deem enough for our purposes however when testing on our own Twitter vaccination data set with manual evaluation too many obvious errors are made and therefore rendering the classifier as useless. We suspect that the problem is that the learned classifier on the general tweets could not be transferred to the specific domain of vaccination.

---

<sup>1</sup> [https://www.kaggle.com/vkrahul/twitter-hate-speech?select=test\\_tweets\\_anuFYb8.csv](https://www.kaggle.com/vkrahul/twitter-hate-speech?select=test_tweets_anuFYb8.csv)

---

As an alternative we decided to use a simple approach of lexical based classification. We use the hate speech lexicon of Wiegand et al. (2018) reduce it to the most relevant 1238 and remove some weak words (such as horrible, disgusting and scissor). If a tweet contains a term of the resulting hate speech lexicon, we classify it as hate speech. This approach is more stable across different domains since the hateful vocabular stays the same.

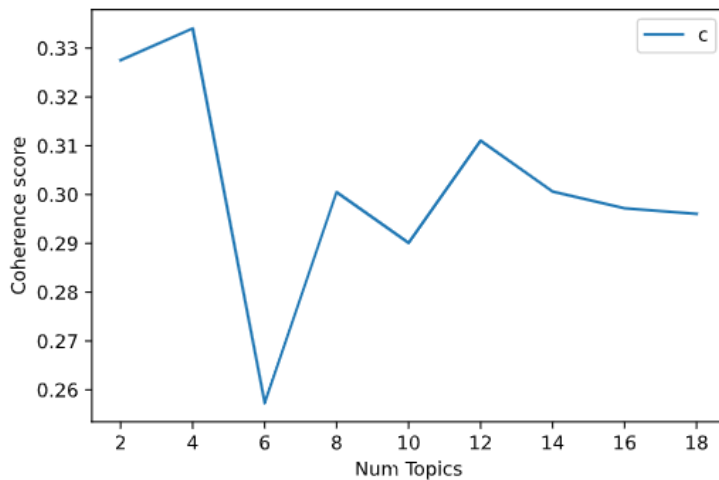
## Topic Modeling

For the topic modeling analysis, we use the Latent Dirichlet Allocation (LDA) based on the lecture's notebook with the genism package. First, we build multiple LDA models with the number of topics ranging from 2 to 20. We then plot the resulting coherence score and choose the final topic number based on inflection points. As the next step we study the most salient terms for each topic to identify a topic description. Finally, we annotate the Tweets with its dominant topic.

Running the analysis for the first time we obtain a suitable topic number of 13. However, when counting the dominant topic for each Tweet we classify 22499 Tweets (99.84 %) with the same topic, which is impractical for further analysis. We assume that because the common topic among all Tweets is COVID and vaccines all those Tweets are dominant with the topic. We repeat the analysis with removing all terms equal to vaccine, vaccination, vax, vaccinated or covid for the Tweets. We choose 12 as the number of topics (see the corresponding coherence score plot below) and find the following phrases and counts:

Topic	Dominant Topic Frequency
Risk/Country/mRNA	11351
Death/Child	6431
Biden/Trump	1771
Death/Infection	824
Mask	781
Passport/Citizen	653
Experiment	226
Blood cloth/Research	211
People get (vaccinated)	116
Administration/Effectiveness	109
Fight/Plan	54
Herd immunity/Republicans/Sharing	7





## Named Entity Recognition

The named entity recognition (NER) is based on the lecture's notebook and uses the spaCy NER. We extract all persons, nationalities, religious or political groups (NORP), organizations (ORG) and geopolitical entities (GPE).

The five most frequent entities of all categories can be seen in the tables below. Some are wrongly labeled, although most are accurate.

Person	Frequency
biden	1019
vaccine	771
trump	598
pfizer	556
moderna	259
joe	259

NORP	Frequency
americans	529
american	323
chinese	120
republicans	90
covid	90

ORG	Frequency
cdc	506
fda	385
trump	307
j&j	128
learn	92

---

GPE	Frequency
us	1482
g7	995
india	499
america	365
uk	299

# Combined Analysis

Finally, we will analyze the Tweet annotations in combination to gain additional insights regarding the problem. In the analysis we will calculate multiple percentages of Tweet classification given a Tweet satisfies a relevant condition. The calculation for class  $c$  given condition  $e$  will be made like the following:

$$f(c | e) = \frac{|\{t \in T \mid \text{class}(t)=c, \text{satisfies}(t,e)\}|}{|\{t \in T \mid \text{satisfies}(t,e)\}|} \quad \text{with } T \text{ being the set of all Tweets}$$

We calculate this for the classifications sentiment and hate speech and for conditions of topics, mentions, NORPs, ORGs and GPEs. The following table shows the values for the classification of sentiment and hate speech on the condition of topic, the remaining values can be found in the corresponding notebook.

Topic	Sentiment			Hate speech	
	Positive	Negative	Neutral	True	False
Death/Child	0.533199	0.299020	0.167781	0.062665	0.937335
Risk/Country/mRNA	0.434852	0.383490	0.181658	0.092943	0.907057
Mask	0.997439	0.000000	0.002561	0.000000	1.000000
People get (vaccinated)	0.551724	0.198276	0.250000	0.060345	0.939655
Death/Infection	0.391990	0.467233	0.140777	0.095874	0.904126
Experiment	0.548673	0.283186	0.168142	0.066372	0.933628
Blood cloth/Research	0.748815	0.208531	0.042654	0.080569	0.919431
Passport/Citizen	0.675345	0.284839	0.039816	0.013783	0.986217
Administration/Effectiveness	0.467890	0.321101	0.211009	0.036697	0.963303
Biden/Trump	0.469791	0.373800	0.156409	0.086957	0.913043
Fight/Plan	0.481481	0.259259	0.259259	0.018519	0.981481
Herd immunity/Republicans/Sharing	0.571429	0.285714	0.142857	0.142857	0.857143
All	0.495562	0.341306	0.163131	0.077439	0.922561

# Conclusions

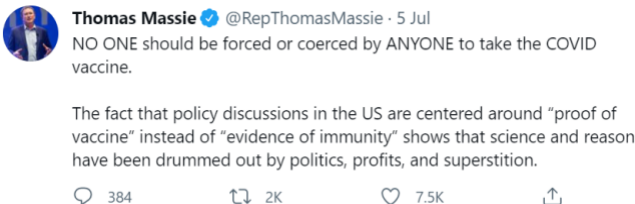
As marketing managers, we looked at all the data extracted, gather the one that could be more relevant to analyze and drive conclusions from it. Only after that we can make suggestions to US Government on how to deal better with public opinion regarding COVID vaccines. Additionally to the following Conclusions and Suggestions we also provide a vastly annotated data set of Tweets that can be used for further analysis.

Looking for all the 23k tweets extracted, we can highlight some accounts that belong to recognized people or institutions due to the amount of retweets and likes they have, making them the main opinion leaders on Twitter regarding Covid topic:

- 1) **@AlexBerenson** with 277k followers, have several tweets with disruptive ideas regarding covid vaccination that are retweeted and liked by thousand of people. He is a former New York Times reporter, Yale-educated novelist, avid tweeter, online essayist and is commonly live at Fox News. On Twitter, he tried to argue that most people would be better if they don't get vaccinated and usually contests the government recommendations regarding covid protection (like wearing masks for example).



- 2) **@RepThomasMassie** with 201,6k followers and many tweets regarding vaccines, his opinion is followed by thousands of people. He is an American Republican Party politician and usually posts tweets regarding government decisions about vaccination.



- 3) **@WHCOS** that represents Ronald Klain, the White House Chief of Staff, has 466k followers and usually makes retweets supporting government decisions regarding vaccination and is one of the accounts with most liked tweets according to Covid topic.





Checking now the **Sentiment Analysis** (Table 1) to understand the public opinion regarding some politicians and organizations, we can state that:

- 1) President Biden (@potus) have a slightly above average positive sentiment and Vice President Harris even more positive.
- 2) Foreign politicians and governments such as Boris Johnson (UK), Justin Trudeau (Prime Minister of Canada), Scott Morrison (Prime Minister of Australia) and the German government Twitter have a very positive sentiment. On the other hand others such as Emmanuel Macron and the Japanese Prime Minister Office have positive sentiments below average and negative above average.
- 3) The Center of Disease Control and Prevention (CDC) has both in mentions as well as in the entity analysis a negative sentiment with almost no neutral Tweets. While the Food and Drug Administration FDA has in both more positive Tweets but also a below average percentage of neutral Tweets.
- 4) International organizations as the EU and G7 are also noticeably more positive.

Taking into consideration the sentiment regarding the different vaccines (Table 2), we can assume that Astrazeneca and Pfizer are the ones that people talk better about and on contrary, Moderna and J&J are the ones not so well received by the public opinion.

From the Sentiment combined with the identified topics we can see that:

- 1) People tweet a lot about death also related to children however the sentiment and the hate speech rate are surprisingly positive
- 2) There are many Tweets (11351) that are related to the risk of vaccines especially with the mRNA technology
- 3) On the other hand Tweets regarding masks are entirely positive although a much smaller number
- 4) Tweets regarding the government (Topic Administration/Effectiveness) and the Biden/Trump dispute are more negative.
- 5) Tweets containing republicans or democrats have a strongly increased amount of hate speech.

Due to the high number of tweets extracted, to simplify the analysis process we can look at the Top5 Tweets with more retweets (Table 4) and with more likes (Table 3), which we can assume are the ones people agree more with. The overall opinion is the unknown vaccines side effects and uncertainty of the efficiency after taking it are the main reasons for people to claim not taking it.

The most common terms used considering frequency analysis are “vaccine”, “covid”, “get”, “vaccination”, “vaccinated”, “people”, “world”, “trump” and “biden”.



---

# Suggestions

US Government needs to follow the controversy tweets from @AlexBerenson and @RepThomasMassie to contest fake news and misinformation. To do so, they can use the twitter account of Ronald Klain or President Biden. However, being official accounts, the more sceptic Americans could not believe on it so, to solve that, the US Government could try to convince Boris Johnson, PM of Canada and PM of Australia to mention that US is doing a good job and work closer on developing vaccines in a faster way to satisfy countries that have no access to the vaccine. That would reinforce the positive feeling regarding US Government. From the frequency and positive sentiment, we can also recommend increasing the partnership with international institutions to increase the acceptance of the administration.

We also numerous Tweets with mostly either positive or negative (so no neutrality) regarding President Biden and former President Trump as well as the Republican and Democratic Party. The administration should try to make a bipartisan cooperation on vaccination promotion initiative to prevent disagreement hindering vaccination progress.

To develop more confidence in the vaccination, US Government should give voice to health specialists to:

- 1) clarify the results from the vaccination;
- 2) why is so important for everyone to get the two doses (or one in J&J) and communicate usage of 3<sup>rd</sup> dose;
- 3) the differences between vaccines;
- 4) what are the side effects of each one according to the age of the patient (especially regarding blood cloth) compared to risks of covid infections;
- 5) what are the long-term side effects children can have from taking vaccine and using mask from such a young age.

Further the US can boost vaccine access in critical states especially among young people since we discovered mentions of difficult access. Users (can be citizens from various countries) show also a lot concern regarding the vaccine availability in other countries. The administration should make plans to tackle global vaccine distribution the vaccination progress in other countries can also affect the national situation.

We believe that with transparency and clarification, people would not have so many arguments to contest the vaccines and would embrace the US vaccination plan because they would realize that would be the only way to end covid spread and have a more normal life like we had 2 years ago.

# Bibliography

- Devlin, J. a.-W. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Minneapolis, Minnesota: Association for Computational Linguistics.
- Wiegand, M. a. (2018). Inducing a Lexicon of Abusive Words - a Feature-Based Approach. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1046–1056). New Orleans, Louisiana: Association for Computational Linguistics.
- Robert-Koch-Institut. (2021). *Epidemiologisches Bulletin* 27. Berlin: Robert Koch-Institut.

## Annexes

	Positive	Negative	Neutral	Nr. Of Tweets
<b>potus</b>	0.512744	0.338646	0.148610	12987.0
<b>cdcgov</b>	0.460184	0.352473	0.187343	4772.0
<b>borisjohnson</b>	0.794913	0.171701	0.033386	1258.0
<b>justintrudeau</b>	0.800976	0.178537	0.020488	1025.0
<b>us_fda</b>	0.514689	0.318449	0.166863	851.0
<b>regsprecher</b>	0.997622	0.002378	0.000000	841.0
<b>scottmorrisonmp</b>	0.968059	0.011057	0.020885	814.0
<b>vp</b>	0.537608	0.334155	0.128237	811.0
<b>eucouncil</b>	0.994792	0.000000	0.005208	768.0
<b>cdcdirector</b>	0.484709	0.391437	0.123853	654.0
<b>kamalaharris</b>	0.591760	0.258427	0.149813	267.0
<b>jpn_pmo</b>	0.145374	0.722467	0.132159	227.0
<b>g7</b>	0.625000	0.260000	0.115000	200.0
<b>speakerpelosi</b>	0.425287	0.408046	0.166667	174.0
<b>erictopol</b>	0.490385	0.288462	0.221154	104.0

**Table 1:** Sentiment Analysis of governments and politicians


	Positive	Negative	Neutral
<b>moderna</b>	0.481013	0.240506	0.278481
<b>j&amp;j</b>	0.425414	0.276243	0.298343
<b>astrazeneca</b>	0.578313	0.228916	0.192771
<b>pfizer</b>	0.567442	0.204651	0.227907

**Table 2:** Sentiment Analysis of vaccines



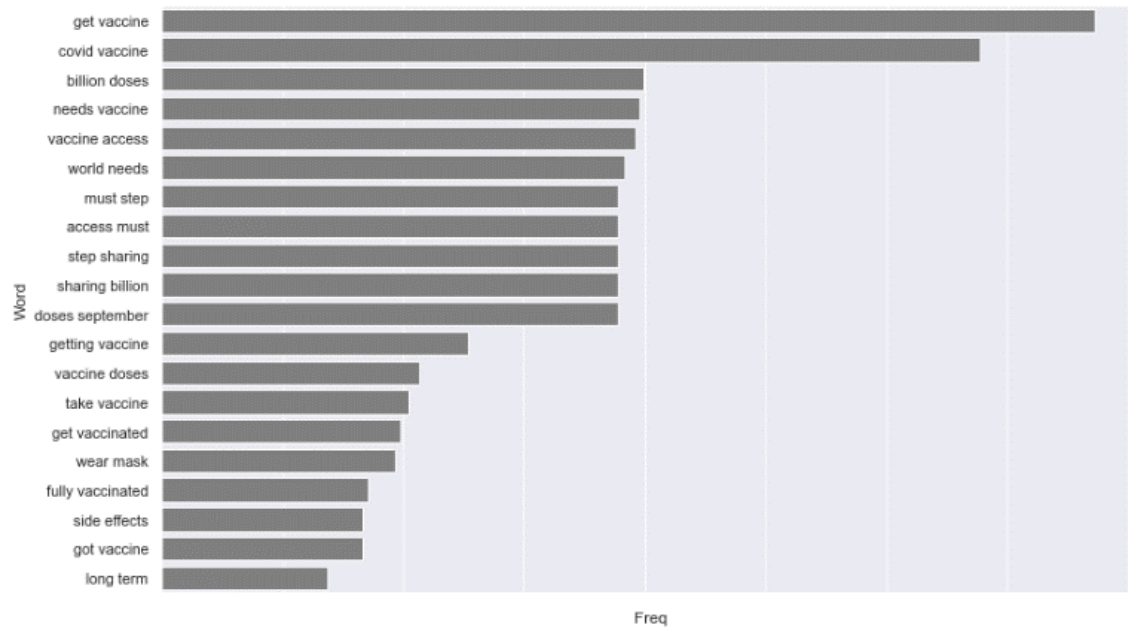
text	user screen name	user followers	retweets	likes	Sentiment	Polarity
I'm not advocating anyone should seek to get infected with COVID instead of taking the vaccine. But I find it appalling that , , & are denying science and treating people like idiots by refusing to acknowledge the efficacy of a natural immune response.	RepThomasMassie	186207	555	3296	-0,4019	Negative
I/ has now analyzed the VAERS data on Covid vaccine myocarditis in teens and young adults. It is terrible.  Based only on received reports - and remember, most side effects go unreported even when they are serious - the rate is as high as 40 times the background rate...	AlexBerenson	267619	1402	2285	-0,5267	Negative
I've been saying this since last year. I even showed the vaccine trial data showed this. They privately acknowledged I was right, but then continued to lie about what the trial data showed. covered their cover-up months ago.	RepThomasMassie	189297	849	2070	-0,4215	Negative
Been a part of several reopening convos in the last 48 hrs, from small restaurants to big multinationals  I common theme: confusion and the risk for many different approaches instead of just one  needs to rip off the bandaid and (at least) recommend vaccine cert for no mask	VinGuptaMD	95721	253	1642	-0,7657	Negative
Those jobs aren't being created, they are just merely coming back now that more people get the vaccine	RickTemple14	215	24	1489	-0,1877	Negative
Why are so many Dems confused "suddenly" by 's CDC...anyone thats wants a vaccine can get one, so far the vaccine works - cases plummeting...we can take mask off indoors/outdoors, game over...you have freedom..Go back to your life!!	kilmeade	490122	172	1481	-0,485	Negative
PARENTS: the covid vaccine fanatics are coming for your kids at school. They aren't even hiding it. "School-focused strategies" start next month.  And they as much as tell schools not to ask for parental consent unless required: these are "routine immunizations," says.	AlexBerenson	267619	660	1136	-0,0772	Negative

**Table 3: TOP5 Tweets ordered by number of likes**

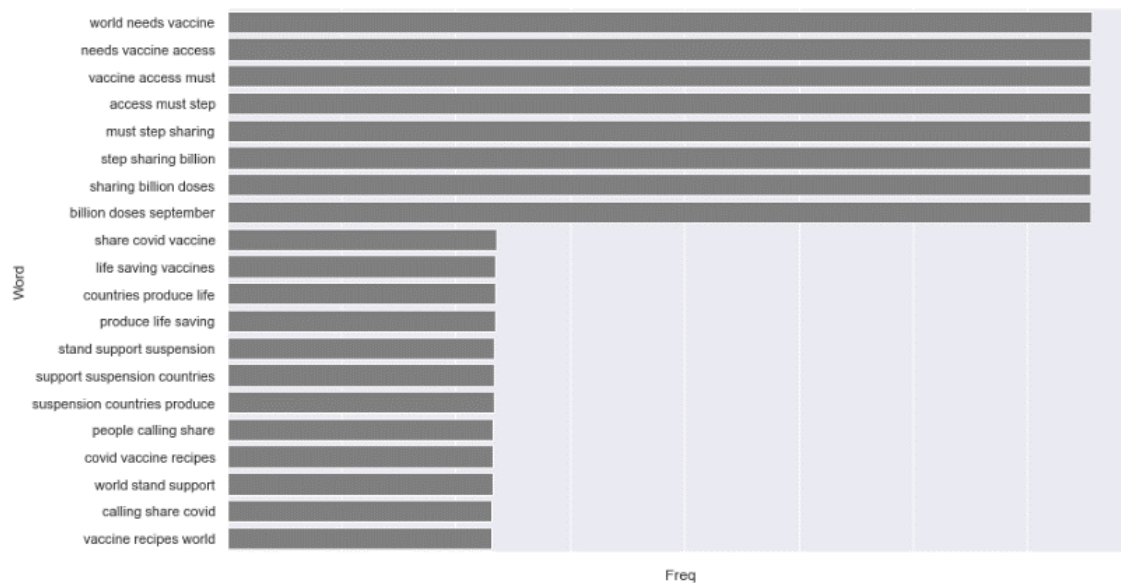
text	user screen name	user followers	retweets	likes	Sentiment	Polarity
I/ has now analyzed the VAERS data on Covid vaccine myocarditis in teens and young adults. It is terrible.  Based only on received reports - and remember, most side effects go unreported even when they are serious - the rate is as high as 40 times the background rate...	AlexBerenson	267619	1402	2285	-0,5267	Negative
I've been saying this since last year. I even showed the vaccine trial data showed this. They privately acknowledged I was right, but then continued to lie about what the trial data showed. covered their cover-up months ago.	RepThomasMassie	189297	849	2070	-0,4215	Negative
PARENTS: the covid vaccine fanatics are coming for your kids at school. They aren't even hiding it. "School-focused strategies" start next month.  And they as much as tell schools not to ask for parental consent unless required: these are "routine immunizations," says.	AlexBerenson	267619	660	1136	-0,0772	Negative
I'm not advocating anyone should seek to get infected with COVID instead of taking the vaccine. But I find it appalling that , , & are denying science and treating people like idiots by refusing to acknowledge the efficacy of a natural immune response.	RepThomasMassie	186207	555	3296	-0,4019	Negative
My unit has been seeing ~2-3 admits/week w/post pfizer/moderna myocarditis in young patients (&lt;45) w/those &lt;30  critical Sx  Cardiologists & hospitalists who round are freaked out. As are the	gebrinson	538	435	668	-0,5423	Negative

nurses.						
Hospital system is now capturing data on anyone coming in post vaccination						

**Table 4:** TOP5 Tweets ordered by number of retweets



**Figure 1:** Tweets Bi-gram



**Figure 2:** Tweets Tri-gram