University of Copenhagen
Department of Economics

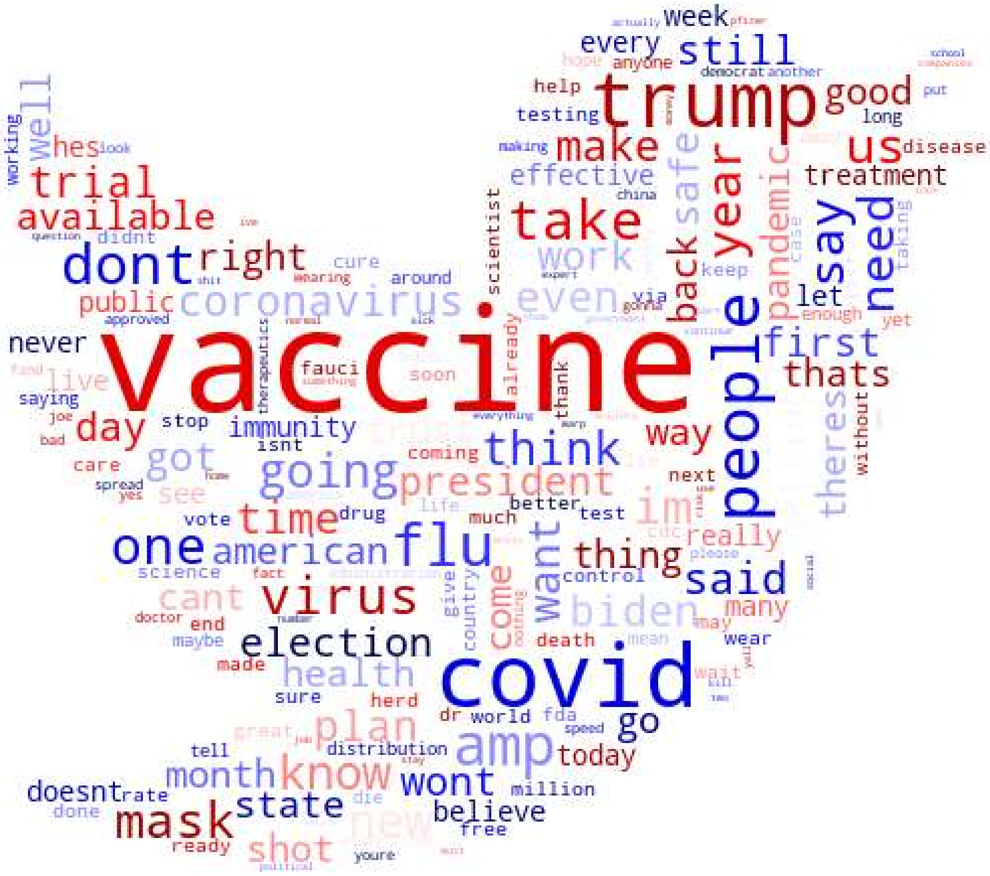
US vaccine sentiment analysis using tweets collected from Twitter

Introduction To Social Data Science

Christian Nøjgaard Fogtdal (rxw556) & Anton Maach-Møller (hfp710)

Date: 22-08-2022
Keystrokes: 26.449
ECTS credits: 7.5

Division: Anton has written the first paragraph and from there.
Christian has written the second paragraph and from there every other paragraph
The Introduction, Discussion and Conclusion have been written in collaboration


vaccine
trump
covid
people
flu
dont
take
year
say
need
virus
one
think
going
coronavirus
even
work
back
safe
pandemic
first
thats
still
good
us
make
trial
available
right
day
way
time
american
election
mask
know
plan
amp
health
biden
want
said
thing
president
got
im
theres
come
cant
go
wont
month
state
doesnt
shot
new
believe
today
immunity
effective
treatment
well
hes
public
never
live
fauci
soon
week
every
help
testing
disease
long
let
via
keep
cure
around
scientist
therapeutics
already
thank
without
really
many
said
control
science
vote
drug
much
better
test
next
wear
death
herd
world
fda
dr
distribution
tell
sure
made
end
doctor
maybe
free
million
rate
done
ready
youre
care
spread
yes
saying
joe
bad
stop
isnt
coming
see
approved
wearing
didnt
question
working
look
china
making
anyone
democrat
another
actually
pfizer
hope
gonna
country
give
mean
may
wait
yet
enough
fact
great
life
die
put
school
case
taking

Contents

1 Introduction **1**

2 Literature review **2**

3 Ethics **3**

3.1 Ethics in Twitter Data 3

4 Data **4**

4.1 Data collection through API-connection to Twitter 4

4.2 Data cleaning 5

5 Methods **8**

5.1 Sentiment Analysis 8

5.2 Bag of Words 8

5.3 Machine Learning 9

6 Results **10**

6.1 Descriptive analysis 10

6.2 Prediction of political wing by classification of tweet 13

7 Discussion **14**

7.1 Ethical considerations 14

7.2 Data and Methods 15

7.3 Enhancing the predictions 16

8 Conclusion **17**

9 Literature **18**

1 Introduction

With over 200 million monthly users[1], Twitter is one of the world's biggest social media platforms and thereby provides a huge amount of social information. Thus, Twitter is a highly popular website for social data science. Since the platform is free to use, the bulk of posts are available to the public, and researchers can quickly gather a huge number of tweets using its Application Programming Interface (API), Webb et al. (2017). Twitter plays and has played a huge role in the political debate in the US in the past year and especially during the Covid-19 period when the vaccine was highly debated[2]. This paper seeks to analyze the public's view on the Covid-19 vaccine from tweets collected on Twitter and thereby sentiment on the vaccine during the 2020 US presidential election. Furthermore, the paper predicts which political wing the tweets originate from given the political orientation of each state.

The data used in the paper is extracted through access from Twitter's API, an open-source location API, Nominatim and a web scrape of Wikipedia. These tweets are analyzed by applying a sentiment analysis tool from Python known as VADER, where the sentiment related to each tweet regarding the Covid-19 vaccine are considered. After the sentiment of each tweet are analysed, the tweets are categorized geographically using API connection to Nominatim. Following all letters are set to lower case, all mentions are removed, and hashtags. With the cleaned data the paper compared each state's sentiment score and after that, we use the Supervised Machine Learning method to predict if it's feasible to classify tweets to a political wing based on the text of the tweets. The paper concludes that there are a small difference in sentiment of the Covid-19 vaccine given the states political stance.

The outline of the paper is dived into the following sections; First, we present the

[1]https://www.searchenginejournal.com/social-media/biggest-social-media-sites/close

[2]https://www.pewresearch.org/politics/2022/06/16/politics-on-twitter-one-third-of-tweets-from-u-s-adults-are-political/

historical literature findings within the same subject. Next, we discuss the ethics of collecting data from Twitter and scraping websites. Section 4 presents the data used and the process of collecting the necessary data to perform the analysis. Next the methods used to conduct the analysis are outlined. Section 6 presents the results and graphical analysis of the sentiments. This is followed by a discussion of the results, methods used, and other considerations. Lastly, the paper concludes the findings.

2 Literature review

In the past years, the focus on social media has risen and literature suggests that the focus on social media as data has become a bigger part of research C. Steinert-Threlkeld (2018). Social media is no more just a network but a place where politicians and citizens can express their opinions and statements on a specific topic. These statements can now be collected to perform studies on a given topic. These topics vary across social science Apoorv Agarwal et al. (2011) covers a Sentiment Analysis of Twitter Data, where they examine the nature of microblogs on which people post real-time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. Brendan O'Connor et al. (2010) use the same microblogs to examine the link between tweets' sentiment to predict the presidential election in 2012. These findings are supported by Ramteke et al. (2016) who they use machine learning to predict the election results of the presidential vote in 2016.

Sentiment analysis is a well-known tool when considering public opinion on a given topic. Chinnasamy et al. (2022) state that sentiment analysis is used to determine user perceptions based on information such as written opinions e.g. Twitter data. The analysis was carried out based on the written opinion, according to the two opinions. Due to the development of the digital age, people often express and publish their thoughts on social media, which makes us unavoidable.

This paper is in line with Chinnasamy et al. (2022) approach on how to handle and analyse Twitter data. This project identify the sentiment on each tweet when the hashtags regarding the Covid-19 vaccine is used and thereby given it a sentiment score. However this paper differs in the way that we analyse the sentiment across the states in the US and use it to predict the presidential election in 2020.

3 Ethics

In the following section, we will consider the ethical whereabouts regarding social data and the data process of collecting data from Twitter. This must be considered due to the increasing focus on privacy and especially on GDPR[3], and the increased focus on data-security and personal information. Past research papers rely on Salganik's (2017) statements on social research in a digital age. These findings and recommendations are useful guidelines when handling data from Twitter.

3.1 Ethics in Twitter Data

The first ethical consideration presented by Salganik (2017) is respect for persons. This is meant by the autonomy of individuals and the protection of individuals with diminished autonomy. In general, researchers should avoid exploiting peoples privacy without their acceptance. When we collect the data necessary we accepted twitters terms of use from their API website. Which means that we are allowed to collect the individuals data needed for this paper. From the Twitter Developer Agreement and policy, we have accepted not to share any personal information and therefore our data presented in this paper have been handled, so any data cannot be traced back to an individual level.

In addition to respect for persons and their privacy, respect for beneficence is one more aspect to consider. This means that the research should balance no harm and benefits from the given research. When the Respect of Persons are complied and we choose not to share personal information, it will be difficult for others to replicate

[3]https://www.datatilsynet.dk/hvad-siger-reglerne/lovgivning

our findings. This example is one perspective of the balance between risk and harm.

From the benefit principle, the Justice principle ensures that not one group is carrying all the risks attached to the research. This specifies that more groups are taking the same share of risk. Thus, this paper treats each group equally and independently. Thereby, the results presented later on have not been manipulated with respect to groups or segmentation. This process and how we collect and handle the data will later be described in the Data section.

Law and public interest must be hold in research papers. Where the beneficence perspective focus on the individual, this principle focuses on the public and it's laws. When collecting data from Twitters API these laws are hold. Nevertheless, when scraping Wikipedia, these laws should be taken into account. Good practice when scraping is to tell the website who you are and what purpose the scraped data should be used to. These considerations have been upheld for this paper.

4 Data

This section outlines our process of collecting data from Twitter by extracting both tweet information and user information. Location data from Twitter is spurious, why we use Nominatim to translate Twitter's location data to readable data. We also scrape election results from the 2020 US presidential election from Wikipedia. This is followed by a description of the cleaning process of the data and provides a general overview of the data used in the analysis.

4.1 Data collection through API-connection to Twitter

Our data consists of tweets from Twitter concerning the Covid-19 vaccine. The data is collected using Twitter's API spanning from the period 2020-09-15 to 2020-11-02. This specific period is chosen since it correlates with the 2020 US presidential election results, which is in our interest.
Twitter has a rate limit of 300 tweets per 15 minutes, which is equivalent to 1 tweet

per 3 seconds. The initial query of the data extraction is therefore of high importance to maintain the highest amount of retention of usable data going through the data cleaning procedure.

Our query parameters are:

- Includes any of the words/hashtags e.g #vaccine, #vaccines, #covidvaccine.
- Do not request retweets
- Language: English (automatically detected by Twitter)
- Has geographical data (on tweets)
- Geographical data is in USA (on tweets)

Our extraction loops 300 requests for each day in the time period. Using the most trending words/hashtags at the time regarding the covid-19 vaccine. However for a full picture, other related hashtag/words are: "vaccines", "covidvaccine", "coronavaccine", "covid19vaccine", "vaccinations", (syringe emoji). Using the words *#vaccine* and *#vaccines*, we are aware that they are not distinct to the Covid-19 vaccine alone. However, due to the massive media attention at the time we extract, we conclude that the vast majority is related to the Covid-19 vaccine. The initial data before cleaning consists of 9.500 tweets.

4.2 Data cleaning

The first variable we clean is the location of the tweets. We want the location to be presented in state, country, so the state location variable is ready to be analysed across states. Afterwards we clean the text of the tweets to be able to analyse the sentiments of the tweets. The results for each state of the 2020 US presidential election is extracted from Wikipedia. And made ready to merge with previous extracted data. Finally the different data sources is merged.

Cleaning of geographical data using API-connection to Nominatim

Before using Nominatim to extract organized location data, we clean the location variable for mentions (@user) and topics (#hashtag).

Geographical data extracted from Twitter can either be tweet-based or user based. Both the user location and the tweet location can be automatically inputted using GPS tracker or manually input in a free text.
Our analysis is based on presidential votes in the USA. The user location would be expected to be the user's home, which is the place the user probably would be voting from. The tweet location is instead the location the user has tweeted the current tweet from, which not necessarily is the user's home. The user's location seems therefore like the best indicator of the user's home. However, after inspecting the data, the user location has a much lower retention rate than the tweet location after the data cleaning. This is probably due to the fact, that user locations are more likely to be inputted manually in free text, which includes spelling mistakes and non-existing places like: "SHE is somewhere under the sea" or "Where Tupac Got Shot". Furthermore, the Twitter query only has filters for the location of tweets and not users. To maintain high retention we choose to use the tweet location instead of the user location.

The reason we need Nominatim to translate location data from Twitter is that the location output is spurious. The location data outputs a single string, and can be in multiple forms, - the most common forms being: ((City, State), (City, Country), (State, Country), (Country), ('Free text')). Nominatim uses OpenStreetMap data to decipher spurious locations to organize data.
Nominatim has a rate limit of 1 request per second, which is another reason we want a high retention rate from the query.

Using Nominatim we extract the variables 'City', 'State', and 'Country' from the spurious location string.
After the desired variables are extracted, we replace empty values with np.NaN for easier handling of data. We remove countries not equal to United States. Even though we input in the query, that we only wanted United States as country, empty values is presented as "None" from Twitter. According to Nominatim "None" is a city in Piemonte, Italy and are therefore removed. Another example is a tweet with

location "The City", which is interpreted as The City of London by Nominatim. We also remove tweets with empty values of state, since this is the core of our analysis.

Scraping 2020 US presidential election results from Wikipedia

By investegating Wikipedias website[4], we find the desired table to scrape for the US 2020 presidential election. After the scrape, we see that a doubleheader occurs. By using the command f-string formatting we put the two headers into one. Subsequently, the columns which contain unwanted information are removed. This is done since the two major parties in the USA are the Democrats and the Republicans and these two are the foucs point for this paper. By investigating the rows we see that Nebraska and Maine have additional rows. They allow their electoral votes to be split between candidates[5][6] and are therefore removed. Lastly we set if a states major votes are either Democratic or Republican. This is done due to we want to merges our Wikipedia data with the Twitter data, and thereby the tweet can be clarified if the state is either Democratic or Republican.

Pre-processing of the tweet text

The text of the tweet is now pre-processed before the sentiment analysis is applied. We replace newline (\n) with white space. We lower case all letters to make sure that similar words are read equally, no matter where in a sentence they are used. We remove mentions (@user) and links starting with *http*, since we are not interested in those. We remove the hashtag sign #, and keep the word itself. We remove digits. We remove links starting with *http* finally we remove non-alphanumeric characters. This includes emoticons. Emoticons can carry a lot of information, however our sentiment analyser, Vader, cannot handle emoticons.

[4]https://en.wikipedia.org/wiki/$2020_United_States_residential_election$

[5]https://legislature.maine.gov/statutes/21-A/title21-Asec802.html

[6]https://nebraskalegislature.gov/laws/statutes.php?statute=32-1038

5 Methods

In the following section we will describe which methods that have applied to perform the analysis. The goal of this section is to determine the method used to investigate the US-citizens perspective on the Covid-19 vaccine.

5.1 Sentiment Analysis

To determine the sentiment of a given tweet we have used the VaderSentiment[7] analysis tool in Python. With this tool we are able to score each tweet text by its sentiment. Vader is rule-based and thereby the analysis is done by mapping the words, where words are categorized as negative, positive and neutral. VADER analyse each word in the sentence and thereby gives the whole sentence a final score. The chosen range of the score varies between -1 and 1, where -1 is a predominantly negative tweet, 1 is positive and 0 is neutral. Due to ethic considerations we have chosen to construct our own examples on a positive and negative tweet.

Vader scores the text below 0.97, which is almost a perfect positive score:

"Thankfully, I have been safely vaccinated for free with the trustworthy Covid-19 vaccine! God bless us all and this great country!"

On the other side of the scale, a predominately negative text is analysed to a sentiment of -0.94 by Vader:

"FAKE NEWS! The flu virus people is at it again... They are lying about the fake illness again!"

5.2 Bag of Words

Bag of Words (BoW) is a simple method to give an overview of the frequency of the words in each tweet. BoW is a method in continuation of sentiment analysis. From our data the amount of words is large and each word has an individual sentiment

[7] https://anaconda.org/conda-forge/vadersentiment

score. The BoW approach thereby defines the most positive and negative words in our data set. However, there are some drawbacks to this approach due to its simplicity. The BoW does not take the whole sentence into account. Thus, there will be some words which not take the exact indicator of the sentiment.[8] This is a problem because some words can have another meaning, dependent on the phrase or the use of irony in a sentence. The BoW is necessary for us to conduct since we use the BoW later on in our supervised machine learning model.

5.3 Machine Learning

Our analysis consist of a machine learning process or predict whether a tweet is a Democratic of Republican tweet. In our second part of the analysis we implement a simple machine learning based on logistic regression from S. Raschka and V. Mirjalili (2017), called Supervised Machine Learning. Applying the logistic regression, which is a portion of a straightforward logarithm derived from the probability ratio of a certain event or binary incidence, we can thereby categorize binary or multi-categorical data. Since we want to predict which party the tweet comes from we can first state that $\theta = \frac{1}{1-\epsilon^{-s}}$ where s is at linear combination of sample features and the weights. With other words this can be describes as the probaility of a given tweet/sample be the property of either a republican or democratic state. In the Supervised machine learning some fundamental concepts exists e.g overfitting and underfitting the model. If one of the stated occurs it will result in inaccurate predictions. By introducing parameter regulations these issues taken into account. To predict the tweets, we implement the supervised machine learning in Python. This is done by the LogisticRegression command and the SKLearn package, the results will be presented in the results section.

[8] https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/

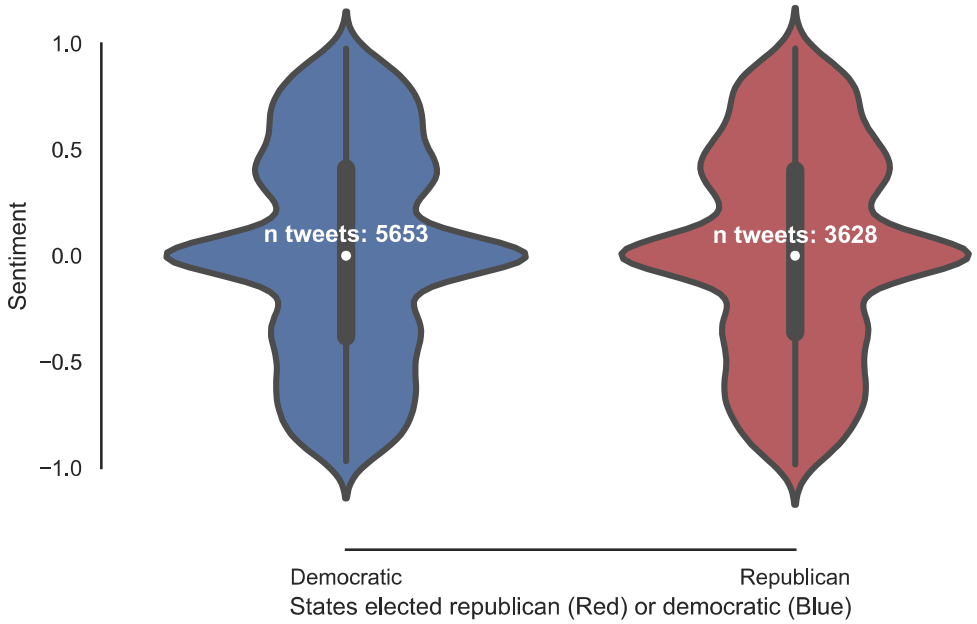
6 Results

In this section, we present the descriptive analysis and prediction results conducted from our machine learning model. First, we focus on the differences in the sentiment score between the republican and the democratic states. Furthermore, we present the sentiment of the Covid-19 vaccine across states to tell the difference at the state level. Secondly, we will focus on the prediction of political orientation based on the tweets regarding vaccines.

6.1 Descriptive analysis

After having cleaned and pre-processed our data collected from Twitter and Wikipedia we are interested in how our cleaned data are presented. From Figure 1, we can see the distribution of the sentiment score divided if the state is either democratic or republican.

Figure 1: Violinplot of sentiment score given political opinion

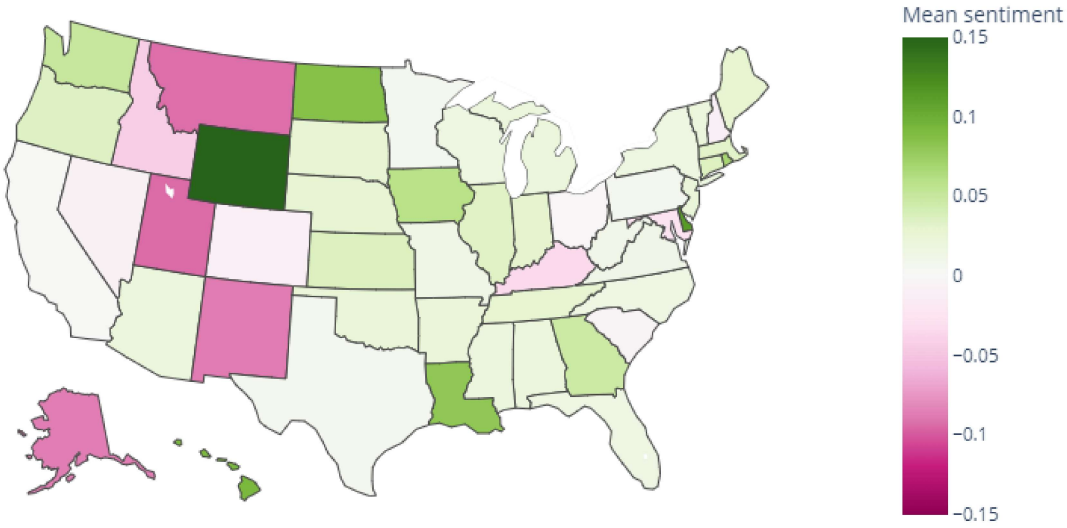


Source: Own Calculations

At first sight, the two violin plots look fairly similar. The number of observations for the democrats is 5,653 tweets sampled and for the Republicans, it is 3,628. Furthermore, the mean for the Democratic states takes a value of 0.015 and for the Republican states takes a value of 0.010. This indicates that the Democratic states are slightly more positive regarding the Covid-19 vaccines. However, by observing the violin plot the distribution is fairly equal and thereby the mean is almost identical. This will be discussed in section 7.

After having seen how the general sentiment is given political orientation, we now want to see the differences across states. This is presented in Figure 2.

Figure 2: Heatmap



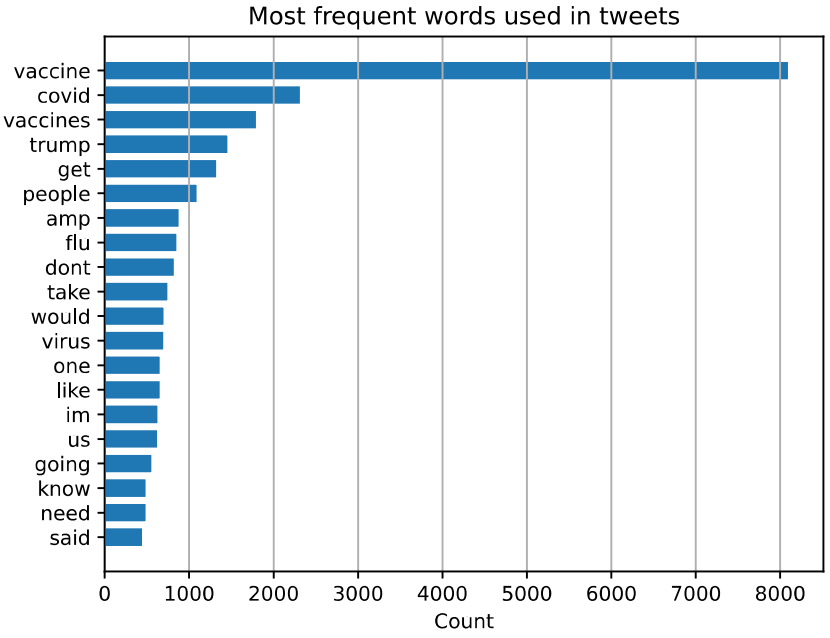
Source: Own Calculations

From figure 2, it clear that the state with the most positive sentiment score is Wyoming with a mean sentiment score taking a value of 0.19, followed by Delaware (0.11) and North Dakota (0.09). On the other hand states as Utah are fairly negative having a mean sentiment score at -0.09. The same applies for Montana (-0.09) and New Mexico (-0.09).
However, it is important to mention, that the number of observations for these states in particular are very low, and thereby probably suffer from having a high variance.

The number of observations and their mean sentiment for each state can be seen in table 3 in appendix.

In Figure 3 we have presented the 20 most frequent words from the cleaned data-set.

Figure 3: Word frequency



Source: Own Calculations

Count from a total of 9,282 tweets

The most and third most frequent word used is 'vaccine' and 'vaccines' which are no big surprise since these are the keyword in our query when extracting data from Twitter. The second is 'covid' and is quite correlated with the word vaccine. The fourth word, trump, is less expected, since the direct correlation between vaccine and trump is hard to explain. However, due to the time period we have chosen, this makes perfect sense, since we have chosen a politically used topic during the US presidential election.

Using the term frequency–inverse document frequency (tf-idf) method, which down-weighs very common and very rare words, we find the most positively and negatively impacting words from our tweets. This is presented in table **??**.

Table 1: 15 most important (positive and negative) features from the tweets

Positive words	value	Negative words	value
great	3.73	flu	−5.18
free	3.61	die	−3.89
safe	3.23	hell	−3.80
effective	2.92	ill	−3.79
better	2.82	shit	−3.76
vaccine	2.79	death	−3.74
safety	2.64	died	−3.65
best	2.43	damn	−3.40
well	2.19	kill	−3.28
good	2.13	bad	−3.23
thank	1.93	fake	−3.08
win	1.92	lies	−3.04
please	1.92	dead	−3.01
help	1.91	lying	−2.96
like	1.90	fuck	−2.95

From the table 1 we see that the most frequent word 'vaccine' gets a positive value of 2.79. Furthermore, the table shows that the 3 most positive word are great, free and safe. The most influential negative words are flu, die and hell.

6.2 Prediction of political wing by classification of tweet

We have now analysed various trends and patterns in the sentiment of tweets regarding the covid-19 vaccine from the US. We now wish to implement a machine learning model, which classifies whether a tweet is originating from either the democratic or republican political wing.

Our model is implemented as a logistic regression as we have described in section 5.3. We divide our data of 9,281 tweets into a training set (70 pct.) and a test set (30 pct.). To reduce dimensionality of the feature matrix and for a higher model interpretability, we choose to only include the 250 most important features from the machine learning model. By using the coefficients obtained earlier from training

the model and the optimal penalization parameter, we predict whether the tweets in the test data are from republican or democratic states. The prediction results is shown below in table 2. The total test data set consists of 2,785 observations, and has 37.45 pct. tweets from democratic states.

Table 2: Prediction results of machine learning model

	Democratic tweets	Republican tweets
Test dataset	1,043	1,742
Predicted	398	2,387
Correctly classified	158	1,502
Wrongfully classified	240	885

The model is overall able to classify 59.61 pct. to the correct political wing. The model is thereby slightly better at classifying tweets compared to a random sample. The shortcomings of the prediction and model is discussed in the following section 7.

7 Discussion

In this section, the methods, results, and other considerations will be discussed. Furthermore, the assumptions made in the paper will be reflected.

7.1 Ethical considerations

In section 3 we described what kind of ethical viewpoints must be considered when using data from a website. As mentioned we use Twitter's API for collecting tweets, but there could still be some ethical problems with doing so. First of all the users from Twitter have accepted that their information could be shared and will be shared in the Twitter API. However, it is most likely that people do not know this, since it's possible that the users didn't read all the conditions related to what they have agreed on. In continuation of the stated it could therefore be unethical to present these data, present personal information. However, the analysis conducted does not

present personal information. Thus, is it fair to say that these ethical issues that could occur are dealt with reasonably.

7.2 Data and Methods

Our boolean variable which defines whether a state is democratic or republican could be considered a large assumption, since most states have a relatively close to 50/50 split in votes[9]. Meaning that the assumption that one state is either democratic or republican is quite hard. Since most of the states are almost equal in votes, this assumption will affect the results. In a way that some Republican political tweets will be Democratic on paper. This is an issue when presenting the sentiment score given by the political orientation and we see from our results, that the distribution is almost identical across states.

Another aspect of the data collection is the total sample size. Due to the limitations from Twitter[10] and time limitations we end up with a total of 9.281 tweets after cleaning the data. This small amount of observations can affect the results. Peter Kokol et al. (2022) state that small sample sizes when considering machine learning will be a problem, because, in general, the machine learning recognizing patterns is proportional to the size of the dataset. This means that small dataset will give less accurate machine learning algorithms.

From the table 3 in appendix we see that those states which are the most positive and negative only got at sample size ranging between 3 and 73. If these states are compared to the three states with most observations, we see that New York have a total of 674 observations, California 1247 and Florida 767. This is a clear indication of, that some states are more represented in the final data-set and thereby it doesn't give a representative distribution for each state.

For the sentiment analysis we have chosen the sentiment analysis tool from Vader.

[9]https://edition.cnn.com/election/2020/results/president

[10]https://developer.twitter.com/en/docs/twitter-api/rate-limits

Vader is rule-based and analyses word-for-word. Thereby, the sentiment score of each tweet is not considered based on the context. Thus, irony, sarcasm and jokes are not captured for the final sentiment score. However, this method is time saving and apply the same method for each tweet. Furthermore, this approach is in line with other studies within the same field Yang et al. (2017). To deal with the above problem, future research studies could create their own using supervised machine learning to capture these missing effects. If doing this, a more nuanced sentiment score could be achieved, which could results in a better model prediction.

7.3 Enhancing the predictions

Since Twitter and Nominatim has rate limits to ensure stable connections, the extraction of data takes time. Using data with state information, we see a clear picture, that some states are much more represented than others. Having a relatively low sample size, the variance of the less frequent states rises, and can become outliers as we see in table 3.

Another reason, that predicting Republican or Democratic states apart based on tweets is hard, is due to the fact that most states have a relatively equal split in votes between democrats and republicans. A possible enhancement of the model could therefore be to remove the so-called "swing-states", which are the states with the most equal votes. The difference between democratic and republican tweets would thereby be more polarized, and easier to classify between each other.

We saw on figure 2 and figure 1, that there is a slight difference between states in terms of sentiment in tweets. An extension of our model could be to include the sentiment scores of the tweets as features for the prediction model. However, while seeing the amount of observations rise, we see that the mean of the sentiment converges towards 0. Without a larger sample size, this is still unknown, but would be interesting for future work.

8 Conclusion

Given our assumptions and relatively limited data, we can conclude, that a slight difference between states in the US can be seen when analyzing Twitter sentiment regarding the Covid-19 vaccine. The analysis shows that the Democratic states are somewhat more positive regarding the Covid-19 vaccine compared with to republican states. However, this difference is minor and might converge towards zero with a larger sample size. We can thereby not conclude that a categorical difference is found between the democratic and republican states. We can thereby conclude that the way a state is equal to the democratic or republican is too harsh and the states more or less have the same distribution in the votes.

In addition, the paper has presented and discussed the ability to predict a tweet's political wing. From the supervised machine learning model, we find that the predicted model is overall able to classify 59.61 pct. to the correct political wing. This is not a great prediction accuracy, but slightly better compared to a random sample.

9 Literature

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau (2011), Sentiment Analysis of Twitter Data Apoorv, Department of Computer Science ,Columbia University

Brendan O'Connor et al. (2010),"From tweets to polls: Linking text sentiment to public opinion time series", Conference on weblogs and social media.

Chinnasamy et al. (2022),COVID-19 vaccine sentiment analysis using public opinions on Twitter, Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad

C. Steinert-Threlkeld (2018), Twitter as Data, Cambridge University

Helena Webb et al (2017), The Ethical Challenges of Publishing Twitter Data for Research Dissemination, Troy, NY, USA

J Ramteke, S Shah, D Godhia (2016), Election result prediction using Twitter sentiment analysis, Department of Computer Engineering, Sardar Patel Institute of Technology

Peter Kokol, Marko Kokol, Sašo Zagoranski (2022), Machine learning on small size samples: A synthetic knowledge synthesis, Faculty of Electrical Engineering and Computer Science, University of Maribor

Ramnath Balasubramanyan (2010), From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, Carnegie Mellon University

Salganik. (2017) Bit by bit: Social research in the digital age, Princeton University Press

S. Raschka and V. Mirjalili. (2017), "Python Machine Learning: Machine Learning and Deep Learning with Python", Second edition.

Yang et al. (2019), Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment, University of Hong Kong

Links:

https://www.searchenginejournal.com/social-media/biggest-social-media-sites/close

https://www.pewresearch.org/politics/2022/06/16/politics-on-twitter-one-third-of-tweets-from-u-s-adults-are-political/

https://edition.cnn.com/election/2020/results/president

https://developer.twitter.com/en/docs/twitter-api/rate-limits

https://en.wikipedia.org/wiki/$2020_{U}nited_{S}tates_{p}residential_{e}lection$

https://anaconda.org/conda-forge/vadersentiment

https://legislature.maine.gov/statutes/21-A/title21-Asec802.html

https://nebraskalegislature.gov/laws/statutes.php?statute=32-1038

Appendix

Table 3: States, their amount of tweets and their mean sentiment, part 1

State	State code	Count	Mean sentiment
Alabama	AL	106	0.0202
Alaska	AK	27	-0.086
Arizona	AZ	226	0.0205
Arkansas	AR	55	0.023
California	CA	1247	0.0023
Colorado	CO	147	-0.011
Connecticut	CT	54	0.0361
Delaware	DE	31	0.1135
District of Columbia	DC	77	0.0336
Florida	FL	767	0.0152
Georgia	GA	220	0.0469
Hawaii	HI	35	0.0917
Idaho	ID	40	-0.0431
Illinois	IL	276	0.0339
Indiana	IN	157	0.0297
Iowa	IA	62	0.059
Kansas	KS	107	0.0375
Kentucky	KY	114	-0.0336
Louisiana	LA	66	0.0813
Maine	ME	52	0.0276
Maryland	MD	274	-0.0238
Massachusetts	MA	249	0.0324
Michigan	MI	192	0.0211
Minnesota	MN	186	0.006
Mississippi	MS	50	0.0192
Missouri	MO	154	0.0136
Montana	MT	15	-0.0925

Source: Own calculations i Python

States, their amount of tweets and their mean sentiment, part 2

State	State code	Count	Mean sentiment
Nebraska	NE	49	0.026
Nevada	NV	153	-0.0063
New Hampshire	NH	66	-0.0093
New Jersey	NJ	297	0.0241
New Mexico	NM	51	-0.0875
New York	NY	674	0.0177
North Carolina	NC	268	0.0147
North Dakota	ND	11	0.0862
Ohio	OH	339	-0.0029
Oklahoma	OK	128	0.0243
Oregon	OR	155	0.0344
Pennsylvania	PA	333	0.006
Rhode Island	RI	36	0.0685
South Carolina	SC	99	-0.0041
South Dakota	SD	26	0.0269
Tennessee	TN	197	0.0281
Texas	TX	673	0.006
Utah	UT	73	-0.094
Vermont	VT	30	0.0196
Virginia	VA	240	0.0125
Washington	WA	242	0.0506
West Virginia	WV	42	0.0112
Wisconsin	WI	110	0.017
Wyoming	WY	3	0.1702

Source: Own calculations i Python