A thorough analysis of the misinformation surrounding vaccines on Twitter, and the political leanings of those promoting this misinformation

Nathan Luc Evans

School of Computer Science and Informatics

Cardiff University

MSc Computing

Supervised by Professor Alun Preece

October 2020

Contents

Abstract

This project conducts a thorough analysis of misinformation surrounding vaccinations in an attempt to understand what misinformation is causing people to distrust vaccines. In addition, a unique algorithm to determine whether a user is left or right-wing will be applied to the data so as to analyse the differences in vaccination misinformation between these two groups. The Twitter data had been flagged as “misinformation”, and was provided by the Crime and Security Research Institute. The project is split into two main parts, with the first focussing on the development and validation of the left/right algorithm, which uses word correlations in order find “left and right-wing words”. Meanwhile the second part of the project concerns the analysis of the Twitter data where the subject relates to vaccines. The analysis that was conducted showed conspiracy theories dominated the vaccination dataset, and these were largely promoted by right-wing accounts.

Chapter 1 – Introduction and background

Introduction

In 1998 the medical journal, The Lancet, published a study conducted by the British physician, Andrew Wakefield (1998). The study demonstrated a link between the measles, mumps, rubella vaccine (MMR) and autism. However, the study was found to be fraudulent (BMJ, 2011), with it being flawed both scientifically and ethically. As the BMJ (2011) explains in their article the study relied on parental recall and beliefs which led to the speculative conclusions. Additionally ten out of the twelve co-authors retracted the interpretation of the original data. The retraction stated “no causal link was established between MMR vaccine and autism as the data were insufficient”. Wakefield was found to be guilty of deliberate fraud, as he picked and chose data that suited their case (Rao and Andrade 2011). Also Wakefield failed to disclose serious conflicts of interest when submitting the paper for publication. For instance, he and several of the children in the study were part of ongoing lawsuits attempting to show that there was a link between MMR and autism (NHS, 2010). These developments would cause The Lancet to retract the paper in 2010 (Eggerston, 2010).

In 2010 Wakefield was investigated by the GMC and was struck off the medical register. The panel found that he had acted unethically and was “dishonest and irresponsible” when publishing The Lancet paper (Cooper, 2010).

However, despite the fact that Wakefield was found to be fraudulent and discredited, it left a very damaging legacy. Since the study, a sharp decline in vaccination rates has occurred. As Hussain e*t al.* (2018) writes the MMR vaccination rate dropped from 92% to 84% between 1996 and 2002. There was also a 2% decline in the US. This has led to preventable and previously eradicated diseases making a comeback. The US had eliminated measles in 2000, but now that status is under threat. In 2019 there were more than 1200 measles cases recorded in America, the largest figure since 1992 (Belluz, 2019).

In the scientific community there is no debate regarding the safety and effectiveness of vaccines. Decades of experience and research show that vaccines are effective and safe (Lopez, 2016). Meanwhile, online there clearly is a debate, where as much as half of tweets about vaccination

contain anti-vaccination beliefs (Broniatowski *et al.* 2018). This has led to fears that the anti-vaccine movement could undermine efforts to end the coronavirus pandemic (Ball 2020).

The anti-vaccination movement (often called “anti-vaxxers”) is growing and tech platforms such as Facebook, YouTube and Twitter have amplified their messages and conspiracy theories (Schaffer 2019). As Igoe (2019) writes, “Social media platforms are usually free and accessible, and information is not vetted as with a news source.” There are no checks on what someone posts on social media sites and forums, and this will help facilitate the spread of false information. There is also a study to show that false information spreads considerably faster than true stories. The study was conducted by Vosoughi, Aral and Roy from the Massachusetts Institute of Technology (MIT). They found that false stories travel quicker due to people retweeting false stories, and they are 70% more likely to be retweeted than true stories (Dizikes 2018). Therefore, one could assume that anti-vaxxers are winning the online war.

Aims and Objectives

The aim of this project is to conduct a thorough analysis on vaccine misinformation online to establish what is causing people to distrust vaccines, and whether there are differences in the content depending on whether a user is politically left or right-wing.

In order to meet this aim, the following objectives will need to be fulfilled:

1. **Create a robust algorithm which will determine whether a Twitter account is left or right-wing**.
2. **Be able to reliably validate this algorithm.**
3. **Conduct analysis on vaccination-related data such as n-grams, hashtags use, topic modelling etc.**
4. **Applying the left or right-wing algorithm to the vaccination dataset and comparing the results of the left and right using the analyses in 3.**

Background material

There have been many studies exploring the online anti-vaccination movement and what tactics anti-vaccination authors use. In a study titled *The Anti-vaccination Movement: A Regression in*

Modern Medicine conducted by Hussain, Ali, Ahmed and Hussain (2018), it was found that anti-vaccination authors used dishonest and deceitful tactics such as claiming not to be “anti-vaccine”, but to be “pro-safe vaccines”, whilst also claiming vaccines contained toxins and are unnatural. These tactics were found to be very effective on parents. The study also found that viewing anti-vaccine websites caused an anti-vaccine sentiment which persisted five months later, thereby causing the children of these parents to receive fewer vaccines than recommended.

An article by Science Daily (2020) explains in a study of vaccine knowledge and media use by researchers at the Annenberg Public Policy Center of the University of Pennsylvania, that people who rely on social media for information were likely to be misinformed about vaccines. It was found that up to 20% of respondents were somewhat misinformed about vaccines.

This is emphasised by a survey conducted by the University College London of 70,000 people regarding UK attitudes and behaviour during the coronavirus pandemic. It was found that around 20% of people are likely to refuse a Covid-19 vaccine when it becomes available. Just under half of respondents said they were “very likely” to get vaccinated. The survey also looked at the underlying reasons why someone would be resistant to receiving a vaccine. The results suggested “substantial levels of misinformation amongst the general public about vaccines” (Boseley 2020).

In regards, to whether those promoting anti-vaccine content are more likely to be left or right-wing there are conflicting results. For instance, as McCoy (2017) writes in The Conversation, a website that publishes news stories written by academics and researchers, some researchers find its strongest support in the political left. However, other research finds that conservatives are more likely believe misinformation around vaccines, and that liberals are more likely to endorse pro-vaccination statements. In addition, the more an individual identifies with the Republican Party, the more likely they are to have a negative opinion of vaccination.

The website, Vaccines Today (Finnegan 2019), a member of the WHO-led project Vaccine Safety Net (WHO 2020), has also discussed the issue of political views influencing vaccination rates. They too accepted there have been conflicting results. In America, libertarians (who tend to be on the right of the political spectrum) may object to compulsory vaccination on grounds they oppose

government mandates. Meanwhile those on the left-wing have expressed a mistrust of authorities and industry to justify their anti-vaccination beliefs

It is patently clear that misinformation about vaccines is affecting peoples' attitudes towards them, and with so much misinformation online this is a problem that will likely escalate, which does not bode well when, at the time of writing, we are in the midst of a global pandemic.

In order to decipher the most frequent themes of misinformation on vaccinations, the following series of analyses will be conducted on the data:

- Frequency of vaccination-related tweets
- Hashtag used
- Retweets
- N-grams
- Topic modelling

The studies into anti-vaccination beliefs and political affiliation provided no definite answer. Nonetheless, this study will attempt to compare the left and right-wing's misinformation regarding vaccinations that is produced online. In order to determine whether a user posting about vaccinations is left or right-wing, an algorithm will be developed which will analyse words, phrases and emojis within a user's account. This should indicate where they are on the political spectrum. What is unique about this is, is that no survey or questionnaire has been distributed. The algorithm purely analyses a user's account.

The project will hopefully give an idea of the most frequent themes of vaccine misinformation online, and whether this misinformation is coming from left or right-wing circles.

Technologies and Libraries used

To conduct the necessary analyses of the data, a selection of different technologies will need to be used. The project was conducted exclusively using the programming language Python. Python is a very popular and flexible programming language with a rich set of libraries and tools which are perfect for data science (Srivastava, 2019). The version used was Python 3.6.9 which met all the necessary requirements.

The web-application Jupyter Notebook was used to conduct the necessary analysis of the project. Jupyter Notebook provides an environment where the user can independently run segments of their code, and have those segments displayed to the user. This avoids having to execute the code from the start of the script (Dar 2018).

Jupyter Notebook is very useful as it allows the user to document their code, immediately see the results, data modelling and visualising data such as graphs. Therefore, the presentation of data was something that Jupyter Notebook facilitated very easily.

Data Collection

The Twitter data was mostly provided by the Crime and Security Research Institute (CSRI), and it was collected using a set of search terms focused upon “misinformation”. This set of terms was compiled by a team of social scientists and computer scientists. The aim of these terms was to collect tweets where the author is talking about misinformation, and this will be called out by users e.g. “this is fake news <Link>”. The Twitter data provided covered several different weeks in 2020, from January to July. Twitter data not provided by the CSRI was also collected, using the Python library, tweepy.

Libraries

Data Collection

- tweepy – used for accessing the Twitter API (Tweepy 2020). Used to collect recent tweets from Twitter. Imperative when validating the left/right-wing algorithm.

Text Manipulation

- regex – Regular expressions or regex is a module within Python which provides matching operations on text. Therefore, they are very useful when dealing with a task which involves a lot of text processing (PyMOTW, 2020). In this project they were very useful for stripping text of links, punctuation, numbers etc.

Data Visualisation

- matplotlib – an extensive library used for creating static and animated visualisations in Python (matplotlib, 2020). Very useful for creating graphs and charts.

- seaborn – a Python data visualisation library that is based on matplotlib. Provides informative statistical graphics such as heatmaps (seaborn, 2020).

Data Manipulation

- pandas – a data analysis and manipulation tool which provides easy to use data structures in the form of dataframes, which consisted of columns and rows.

- numpy – a Python library used for working with arrays whilst also having functions in matrices which are a powerful tool for data analysis.

Machine Learning

- Sklean – is a machine learning library for Python. It includes algorithms that support vector machines, which for this project was invaluable.

Natural language Processing

- Nltk – natural language toolkit gives a platform for creating Python programs to work with human language (NLTK, 2020). It was very useful for “Topic Modelling”, as it can filter out useless data using “stopwords”. It also allowed for tokenisation of words, refers to splitting larger amounts of text into smaller lines (Tutorials Point 2020).

Chapter 2- Left/Right Algorithm

Objective of the this chapter

The motivation for the study of this chapter is to attempt to identify someone's position on the political spectrum via their Twitter account, i.e. are they left or right-wing? If successful, this algorithm can be applied to the vaccination dataset where the results can be analysed, and hypothesis can be drawn. Therefore, the objective of this chapter is to develop an algorithm that will determine whether an account is left or right-wing.

Background

Before exploring the the methodology in developing this algorithm, it is very important to understand the concept of a political spectrum. Today, the terms left and right-wing are used as symbolic labels for liberals and conservatives (Andrews 2015). In the broadest terms, left and right-wing describe political positions, ideologies and political parties. Left-wing typically favour the redistribution of wealth (i.e. equality), increased state regulation and the inherent belief that the world can be improved.

Meanwhile right-wing generally support minimal government, are against the redistribution of wealth and have the belief the world is fine as it is. Figure 1.0 gives an idea of the opposing views that are held by the left and right (McCandless & Posavec 2010).

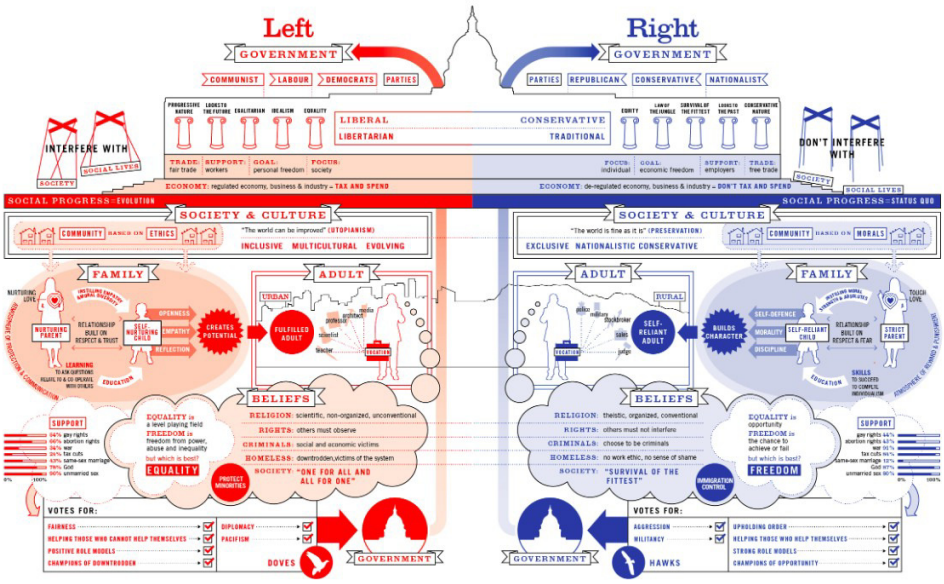


Figure 1.0: Left/right spectrum

Datasets used

To quickly recap, the datasets used in this chapter have been provided by the CSRI, which were collected using a set of search terms focused on misinformation. Since the entirety of the datasets is extremely large, with tens of millions of tweets overall, it is not practical to use all of this data as this will inevitably lead to computer crashes. Instead a sample of around a million tweets will be used which should give a good representation of the entire dataset.

Analytic tools used in the chapter

Below are the analytical tools that were used in this chapter. Please refer to the background section for a more detailed description.

- matplotlib
- nltk
- numpy
- pandas
- regex
- sklearn
- tweepy

5.3 Problem

One of the difficult aspects of the chapter will be to avoid making the left/right algorithm subjective. Words that are deemed to be associated with being “left-wing” should be proven as being left-wing using empirical evidence rather than just observation. One way to do this is to use word correlations, as done by Jatnika *et al.* (2019) who used the Word2Vec Model to turn words from 320,000 Wikipedia articles into vector form and analysed the correlations. This would allow the validation of the word-set as well helping to expand it.

Another problem is how to score the words, phrases, and emojis. Words will be split into strong and weak, with strong obviously scoring higher, and emojis scoring the least. However, there is still the issue of the specific scoring. This problem will discussed in greater detail later on.

Method/Approach

In order to determine whether an account is left or right-wing the analysis will be conducted on the Twitter description of a user i.e. the user-defined UTF -8 string describing their account (Twitter 2020). Certain words relating to political slogans, political parties, ideologies etc. should give an idea of whether that user is left or right-wing.

A user description will be scored on three factors: what words are in the description, the phrases, and finally emojis. A left-wing term will give a negative score, whilst a right-wing term will give a positive score. If the overall score is less than 0 than the account will be regarded as left-wing, whilst if the score is greater than 0 the account will be deemed as right-wing. If the score is equal to 0, the account will be labelled as “Ambiguous”.

Overall score > 0 = Right-Wing
Overall score = 0 = Ambiguous
Overall score < 0 = Left-Wing

To find an adequate set of left and right terms, the words that are being used most frequently in the Twitter description will need to be determined. Firstly, the text in each user description will need to be “cleaned” by removing numbers, punctuation, links, hashtags, capitalisation etc. So-called “stopwords” will also be removed from the text, which are are part of the nltk package. Stopwords are a commonly used words such as “the”, “in” and “a”. These words would take up space in the DataFrame and would not add anything insightful. Please see the figure 1.1 to see words that constitute a stop word.

```
import nltk
nltk.download('punkt')
from nltk.corpus import stopwords
nltk.download('stopwords')
from nltk.tokenize import word_tokenize
```

```
[nltk_data] Downloading package punkt to /home/c1308353/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to
[nltk_data]     /home/c1308353/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
stopwords = nltk.corpus.stopwords.words('english')
print(stopwords)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'y
ourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those',
'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'a
n', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'b
etween', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'of
f', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both',
'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very',
's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'ar
en', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "have
n't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "should
n't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Figure 1.1: stopwords

Another important part of the “cleaning” text process is stemming words. Stemming is the process of reducing inflected words to their word stem, base or root form (Jabeen, 2018). Essentially, words that have the same root meaning will be reduced to the root word. For example see the figure 1.2 (Jabeen, 2018) below. Playing, Plays and Played all share the same root word “Play”, so after stemming the words become “Play”. This is very useful as it it avoids having multiple words with the same meaning in the data.

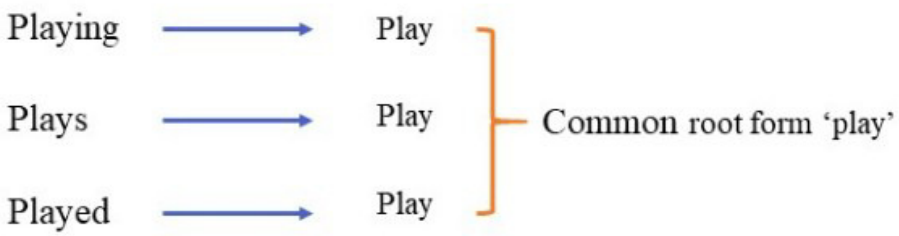


Figure 1.2: Word stemming

```python
def link_remover(text):
    text = re.sub(r"http\S+", "", text)
    text = re.sub(r'bit.ly/\S+', '', text)
    text = text.strip('[link]')
    return text
```

```python
# code used to clean text
# This code was taken from a git hub page by the user James
# Accessed on 5/08/2029
# https://ourcodingclub.github.io/tutorials/topic-modelling-python/

word_rooter = nltk.stem.snowball.PorterStemmer(ignore_stopwords=False).stem
punctuation = '!"$%&\'()*+,-./:;<=>?[\\]^_`{|}~•@'

def clean_text(text, bigrams=False):
    if text is None:
        return ""
    else:
        text = link_remover(text)
        text = text.lower() # lower case
        text = re.sub('['+punctuation + ']+', ' ', text) # strip punctuation
        text = re.sub('\s+', ' ', text) #remove double spacing
        text = re.sub('([0-9]+)', '', text) # remove numbers
        text_token_list = [word for word in text.split(' ')
                           if word not in stopwords] # remove stopwords

        text_token_list = [word_rooter(word) if '#' not in word else word for word in text_token_list] # apply word rooter
        if bigrams:
            text_token_list = text_token_list+[text_token_list[i]+'_'+text_token_list[i+1]
                                               for i in range(len(text_token_list)-1)]
        text = ' '.join(text_token_list)
        return text
```

Figure 1.3: Cleaning the text

The function in figure 1.3 shows the process of cleaning the user description. First, all links are removed, and the casing is put into lower case. The punctuation is stripped and any double spacing is removed along with numbers. Stop words are then removed before the words are stemmed.

An example of a user description going through this function is shown below. At point 1, is the description before the “cleaning”. At point 2 the text is now in lower case, and the punctuation has been removed along with the stop word “in”. The words “freelance”, “photographer”, and “based” have all been stemmed. Then at point 3 any remaining non-alphabetical characters are filtered out using regular expressions and then the words are added into a list.

1. ‘Freelance photographer based in Tokyo. #resist’
2. ‘freelanc photograph base tokyo #resist’
3. [‘freelanc’, ‘photograph’, ‘base’, ‘tokyo’, ‘resist’]

By copying these lists of words into a new DataFrame, flattening it and then grouping the data by what the word is we can see the most frequent words that are used in the user description.

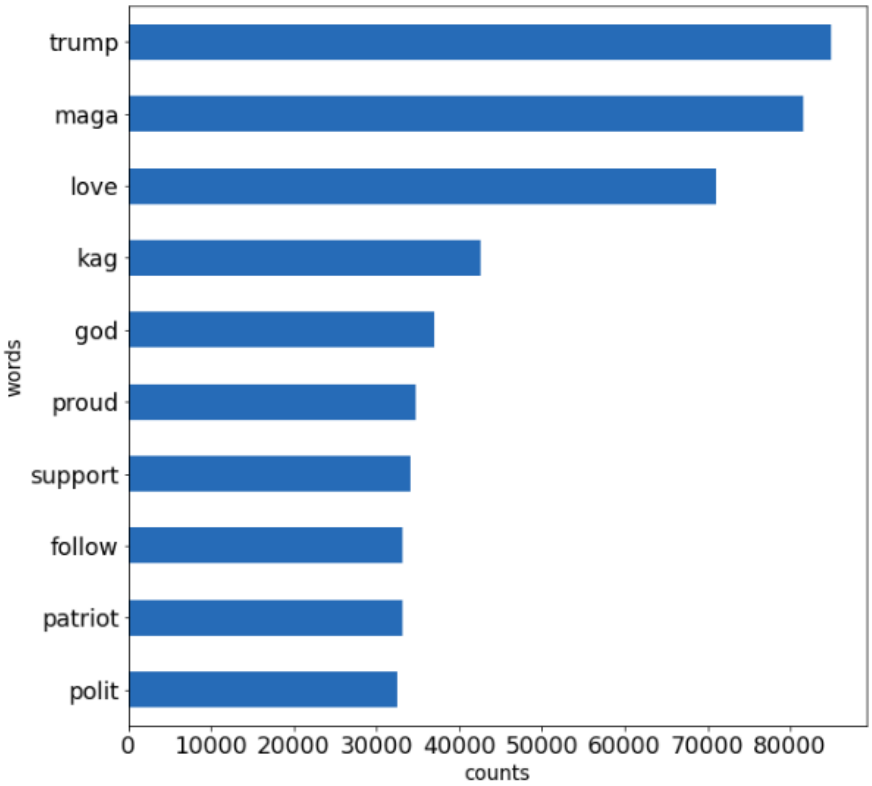


Figure 1.4: Most frequent words in the descriptions

Figure 1.4 shows the results of the most frequently used words in the Twitter description. The results clearly suggest a strong US bias with “trump” and “maga” (Make America Great Again), the political slogan used by Donald Trump, both being most frequently used words.

In order to possess an appropriate number of left and right-wing terms the use of word correlations will prove to be invaluable here as this will help validate the collection of terms by finding expecting neighbours. In order to find out which words are correlated to each other the text will need to be converted to numeric form. For example, if the filtered words from a user description were:

[“maga”, “trump”, “support”] - and the available words in the set were:

[“qanon”, “maga”, “resist” , “world”, “trump”, “support”]

- the vector form of this would be: [0, 1, 0, 0, 1, 1].

Once in vector form, the inbuilt method .corr() can be applied to find correlations between each column of the DataFrame, and therefore a correlation between the different words appearing in the same user description (Coding club 2018). Figure 1.5 shows a snippet of the correlations. The closer the figure is to +1.0 the more correlated those two words are, i.e. they are used together more frequently.

	d	understand	mi	free	hypocrisi	jazz	sex	wrong	lockdown
d	1.000000	-0.002377	-0.001452	-0.005213	-0.001222	-0.001286	-0.001155	0.044035	-0.001567
understand	-0.002377	1.000000	-0.002930	-0.001678	-0.002466	-0.002595	-0.002330	-0.003987	-0.003163
mi	-0.001452	-0.002930	1.000000	-0.006427	-0.001507	-0.001586	-0.001424	-0.002436	-0.001932
free	-0.005213	-0.001678	-0.006427	1.000000	0.007440	0.002449	0.003954	0.004533	-0.000251
hypocrisi	-0.001222	-0.002466	-0.001507	0.007440	1.000000	-0.001335	-0.001198	-0.002050	-0.001626
...	...	...	...	...	...	...	...	...	...
oh	-0.001515	-0.003058	-0.001868	-0.003249	-0.001572	0.012063	-0.001486	-0.002542	-0.002016
heal	-0.001593	-0.003214	-0.001964	0.006111	-0.001653	-0.001739	-0.001562	-0.002671	-0.002119
healer	-0.001307	-0.002637	-0.001611	0.002230	-0.001356	-0.001427	-0.001281	-0.002192	-0.001739
jack	-0.001327	-0.002678	-0.001636	-0.001928	-0.001377	-0.001449	-0.001301	-0.002226	-0.001766
song	-0.001244	-0.002510	-0.001534	0.011330	-0.001291	-0.001358	-0.001220	0.041482	-0.001655

Figure 1.5: Word correlations

Developing the right-wing word-set

Along with showing a strong US bias, there also seems to be a stronger presence of right-wing terms, or at the very least, Donald Trump supporters. The political positions of Donald Trump himself have frequently changed. Since 1987 he has changed party affiliation five times (Gillin 2015). This is emphasised by the findings of the website “On the Issues”, a non-profit organisation providing information for voters. From 2003-2011 they classified him as a “Liberal-leaning populist” (2003). Whilst from 2017, the year he was inaugurated as President, On the Issues (2017) classify him as a “Hard-core conservative”. In February 2017 Trump, in an effort to galvanise his conservative base, described himself as a “nationalist”. Nationalism in America is often associated with white supremacy and the extreme right, which is why past presidents usually avoid this term, preferring to use the safer term “patriot” (Baker 2018).

Since Trump is identifying as a strong conservative, and also a nationalist, then it is fair to say that his supporters are going to be right-wing too. This is supported by the exit polls for the 2016 Presidential Election, where 81% of those who voted for Trump identified their political ideology as conservative (New York Times 2016).

Words which have a political meaning, like “maga” can now be checked for words they correlate with. The algorithm should rely on an equal set of left and right terms in order to make it as fair as possible. The word “maga” was the second most common word in the data, and this word is strongly associated with Donald Trump and his supporters, meaning it a strong right-wing term. The top 20 highest correlations with the word “maga” are listed in figure 1.6. The word “kag”, which stands for “Keep America Great” has the strongest correlation. “Keep America Great” is Trump’s 2020 campaign slogan, and would therefore constitute another strong right-wing term.

The word “wwgwga” (known as wwg1wga- the “1” has been removed from the word) is strongly correlated with “maga”, this word relates to the QAnon conspiracy theory.

According to CBS News (2018), those who subscribe to the QAnon conspiracy theory believe that a person who posts on 4Chan message boards, under the name of “Q”, is a high ranking government official. Q states that the president, Donald Trump, is waging a secret battle against a cabal of Satan-worshipping Democrats, Hollywood celebrities and billionaires that secretly run the world. Q also states that they engage in heinous acts such as paedophilia and harvesting the blood of abused children (Wong 2020). QAnon believers use the term “where we go one, we go all”, which they have misattributed to President Kennedy (which is abbreviated to “WWG1WGA”) as their rallying cry (CBS News 2018).

Figure 1.6 shows that words associated with the QAnon conspiracy theory possess high correlations with “maga” and “trump”. The words “wwgwga” and “qanon” appear in over a third and a quarter of descriptions respectively whenever “maga” is there. The QAnon conspiracy theory is clearly strongly associated with Trump supporters and therefore constitute strong right-wing terms.

```
correlations['maga'].sort_values(ascending=False)[0:20]
```

```
kag                0.449934
trump              0.364275
wwgwga             0.336796
qanon              0.275649
nra                0.231593
patriot            0.231325
a                  0.227744
buildthewall       0.178059
q                  0.169896
draintheswamp      0.164076
conservative       0.162262
prolife            0.161625
christian          0.160651
conserv            0.154436
god                0.149030
americafirst       0.142158
presid             0.133694
walkaway           0.128622
marri              0.124229
usa                0.115430
```

```
correlations['wwgwga'].sort_values(ascending=False)[0:20]
```

```
qanon              0.382285
maga               0.336796
q                  0.261937
kag                0.254424
trump              0.237652
patriot            0.218565
thegreatawakening  0.211515
qarmy              0.170825
greatawakening     0.152365
savethechildren    0.132402
darktolight        0.130632
god                0.129583
godwins            0.124190
nra                0.119539
trusttheplan       0.117537
digitalsoldier     0.115812
a                  0.114400
draintheswamp      0.113373
prolife            0.101525
soldier            0.096979
```

```
correlations['patriot'].sort_values(ascending=False)[0:20]
```

```
maga               0.231325
wwgwga             0.218565
trump              0.166467
kag                0.156638
qanon              0.140571
q                  0.139316
american           0.133294
christian          0.130910
conservative       0.105314
ifb                0.101729
god                0.100895
countri            0.089853
jesu               0.086554
veteran            0.076995
love               0.075674
nra                0.075655
a                  0.075333
americafirst       0.073580
darktolight        0.070692
presid             0.069035
```

```
correlations['buildthewall'].sort_values(ascending=False)[0:20]
```

```
nra                0.210630
americafirst       0.210072
vets               0.201669
draintheswamp      0.191595
maga               0.178059
kag                0.142681
a                  0.141577
trump              0.113425
bluelivesmatter    0.112767
qanon              0.092634
trumptrain         0.083983
deplorable         0.083686
conservative       0.083006
backtheblue        0.079560
ifb                0.076599
wwgwga             0.076576
prolife            0.076198
obamagate          0.062235
fb                 0.061229
constitution       0.059294
```

Figure 1.6: Right-wing word correlations

By using word correlations I was able to create a large list of strong right-wing terms which is shown in figure 1.7. It be should explained that the word “trump” is not this list as this may lead to unreliable results. Although there are many accounts expressing their support for Trump, there are also many accounts expressing their disdain for him too, and these should not be included as right-wing.

There is also a set of weak right-wing terms. These were terms that did have a slight positive correlation with a few words in the strong set but are ultimately too weak to be considered a strong right-wing term. For example, the word “marri”, which is the stemmed word of “marry”, has positive correlations with “maga”, “trump”, “kag” and even “wwgwa” (see figure 1.8). However, the highest of these correlations is only 0.12. In addition, the word “marri” on its own would not suggest right-wing links, but from the correlations it does seem to suggest that a description with “marri” is more likely to be right-wing, and for this reason it will be placed into the weak right-wing set which will be scored lower.

```
strong_right_wing = ['maga', 'kag','wwgwga','q', 'qanon', 'nra', 'americafirst', 'buildthewall',
                     'deplorable','draintheswamp','patriot','conservative','greatawakening','ccot', 'prolife',
                     'bluelivesmatter','nationalist','qarmy','istandwithpresidenttrump','deplor','conserv','republican',
                     'christian','thegreatawakeningworldwide','nation','thegreatawakening','savethechildren',
                     'darktolight','trusttheplan','digitalsoldier','wethepeople','godwins','walkaway','backtheblue']

weak_right_wing = ['constitution','god','potus','pill','potu', 'awakening','jesu','marri','wife','husband','christ']
```

Figure 1.7: Right-wing terms

```
correlations['marri'].sort_values(ascending=False)[0:20]
```

happili	0.552239
maga	0.124229
trump	0.098509
kag	0.096695
christian	0.091423
retir	0.084303
year	0.080386
vet	0.077616
conserv	0.076775
two	0.073139
wonder	0.071517
brat	0.069332
yr	0.068184
cathol	0.065643
kid	0.065442
wwgwga	0.062747
grandchildren	0.060893
mother	0.059738
date	0.059572
potus	0.059081

```
correlations['jesu'].sort_values(ascending=False)[0:20]
```

christ	0.408886
savior	0.214130
lord	0.213087
sinner	0.212604
save	0.125577
said	0.113760
except	0.106179
becom	0.106026
fear	0.101926
constitut	0.101012
bill	0.100905
love	0.094726
q	0.090644
god	0.089635
maga	0.089283
goe	0.088453
order	0.088289
follow	0.088216
way	0.087255
patriot	0.086554

Figure 1.8: Weak right-wing terms

Developing the left-wing word-set

Creating a strong left-wing set proved to be more challenging, as right-wing terms are far more prevalent in the dataset. Another problem is the issue of identifying what constitutes a left-wing term, and as previously mentioned, the data has a strong US bias. There are no political parties in America that are socialist or working class. The US is one of the few industrialised societies that does not have a socialist party at national level. However, the Democratic Party has been more willing to embrace those on the left of the U.S. political spectrum (Morales 2012). There are also left leaning politicians within the Democratic Party, such as Bernie Sanders (who describes himself as a "democratic socialist"), Elizabeth Warren and Alicia Ocasio-Cortez. Therefore, a degree of political relativism may have to be used when creating a set of strong left-wing terms.

The 46th most frequent word in the dataset is "resist". This is in reference to the political movement "The Resistance", which is a group protesting against the presidency of Trump. Although this group does comprise those on the left and right, it was originally started by a group of American liberals so this will be classified as a left leaning term.

Figure 1.9 shows the correlations associated with "resist", with the highest correlation being "fbr", which stands for "follow back resistance", followed by "blm", which is an abbreviation for the activist group "Black Lives Matter". These terms will allow the validation and expansion of the set just as was done for the right-wing set. Figure 1.9 displays the lowest twenty correlations for "blm". Many of these words were part of the strong right-wing set and "blm" possesses a negative correlation with these. Therefore, if the strongest right-wing words have the lowest correlation with "blm", then this word constitutes a strong left-wing term. Figure 2.0 displays the left-wing set.

```
correlations['resist'].sort_values(ascending=False)[0:25]
```

```
fbr                   0.191426
blm                   0.176480
bluewave              0.161852
voteblue              0.140136
eds                   0.127613
vaccinessavelives     0.113562
medtwitter            0.107074
votebluenomatterwho   0.095929
troll                 0.088613
biden                 0.079646
theresistance         0.078675
gop                   0.074233
block                 0.073921
vote                  0.069880
lgbtq                 0.062356
wit                   0.062113
concern               0.057452
liber                 0.056108
evil                  0.053605
resistance            0.053352
impeach               0.052961
blacklivesmatter      0.051804
democrat              0.050987
dem                   0.050585
fascism               0.049187
```

```
correlations['blm'].sort_values(ascending=True)[0:20]
```

```
wwgwga      -0.021453
kag         -0.016691
god         -0.016279
q           -0.013633
truth       -0.013564
countri     -0.013152
christian   -0.012992
qanon       -0.012952
athlet      -0.012194
paddl       -0.011915
believ      -0.011791
freedom     -0.011483
author      -0.011322
interest    -0.010651
scientif    -0.010641
world       -0.010376
presid      -0.010320
patriot     -0.010207
a           -0.010082
maga        -0.010079
```

Figure 1.9: Left-wing correlations

```
strong_left_wing = ['remain', 'leftist','justic','social','progress','resist','vegan','fbr','theresistance','blm',
                    'bluewave','resistance','alli','feminist','lgbtq','lgbt','queer','humanist','blacklivesmatter',
                     'antifa','fascism','progress','green','stopbrexit','pronoun','environmentalist','secular',
                   'votesblue','votebluenomatterwho','fbpe','vegetarian','labour','cannabi','ethic']

weak_left_wing = ['equiti','transgend','gay', 'bisexual','lesbian','hippi','marxist','socialist','anarchist']
```

Figure 2.0: Left-wing terms

Developing left/right-wing phrases

A problem with just looking for words is that it will miss collection of words which could have a political message. For instance if an account has “Make America Great Again!” in their description, this will be reduced to [make, america, great] when the text gets cleaned. This means it will not get picked up when checking for individual words against the right-wing word-set.

Therefore, a function that will look for specific collection of words is needed. Figure 2.1 displays the list of phrases for the left and right.

```
right_wing_phrases = ['drain the swamp','build the wall','america first', 'make america great',
                      'making america great', 'keep america great','trump2020','2a', 'red pill','trump 2020']

left_wing_phrases = ['he/him', 'she/her', 'they/them','she/they','he/they', 'black lives matter', 'human right',
                     'bernie2020','feelthebern','notmeus','followbackresistance']
```

Figure 2.1: Left/right-wing phrases

Emojis

Another function will also check for the use of emojis. Some emojis have had political messages attached to them, so looking at these could give an indicator as to whether an account is left or right.

```
right_wing_emojis = ['🐘','💊','❌','🦅','🐸','👌','✝']

left_wing_emojis = ['🌹','🌈','🏳️‍🌈']
```

Figure 2.2: Left/right-emojis

Emojis associated with the left:

- Rose emoji - symbol of socialism, also the logo of British Labour Party.
- Rainbow emoji – symbol of the LGBT movement.
- Rainbow flag – flag of the LGBT movement

Emojis associated with the right:

- Elephant emoji- symbol of the Republican Party.
- Pill emoji – associated with the red pill and blue pill meme representing a choice between taking a red pill to learn the truth, or taking the blue pill to remain in blissful ignorance. The terms are derived from a scene in the film The Matrix and is commonly used to denote a right-wing political awakening (Swearingen 2020).
- Red cross – Supporters of Donald Trump have recently been using the symbol in their social media handles to express their support of Trump's policies (Dictionary 2020).
- Eagle emoji – national symbol of the US, used to express patriotism.
- Frog emoji – Has been used to represent the online meme "Pepe The Frog". In the early 2010s the meme was appropriated by the alt-right which led to it being declared a "hate symbol" by the Anti-Defamation League (ADL) (BBC 2016).

- Okay hand symbol – Can be used as a symbol for white supremacy, depending on the context. Nonetheless it has been added to a list of hate symbols (BBC 2019).

Emojis will be scored at a very low rate, as it is very possible to use these emojis with no political motive. For example, although the ADL acknowledged the okay hand emoji is being used by some as an expression of white supremacy, they conceded the overwhelming usage of the hand gesture is still to show that someone is okay (BBC 2019).

Scoring System

One of the problems of the scoring system is what numbers to use to score different categories, i.e. how much higher should a strong left/right term score than a weak one? Ultimately, it would be argued that whatever numbers are chosen, it could be subjective. For this reason, it would be perhaps best to show the results using different scoring.

First scoring system:

-/+ 5 for every strong left/right-wing term

-/+ 1 for every weak left/right-wing term

-/+ 5 for every strong left/right-wing phrase

-/+ 1 for every weak left/right-wing phrase

-/+ 0.5 for every left/right emoji

An example:

“WWG1WGA..Blood Bought, God first, my wife and America second. MAGA Trump train passenger. Retired Business owner, Cyrpto curious. I F.B. I am ASPHALTMAN.”

['wwgwga', 'blood', 'bought', 'god', 'first', 'wife', 'america', 'second', 'maga', 'trump', 'train', 'passeng', 'retir', 'busi', 'owner', 'cyrpto', 'curiou', 'f', 'b', 'asphaltman']

words	word_score	Phrase_score	Emoji_score	Overall_score	Political_leanings
[wwgwga, blood, bought, god, first, wife, amer...	12	0.0	0.0	12.0	Right Wing

Figure 2.3: First scoring system

In this user description there are two strong right-wing terms, “wwgwga” and “maga”, and there are two weak right-wing terms “god” and “wife”. There are no left/right-wing phrases or emojis in the example. Using the above scoring, the overall score for this example would be +12, meaning the account would be labelled as right-wing.

Second scoring system: – Strong terms are now 10 times the weighing that of weak terms.
-/+ 10 for every strong left/right-wing term
-/+ 1 for every weak left/right-wing term

-/+ 10 for every strong left/right-wing phrase
-/+ 1 for every weak left/right-wing phrase

-/+ 0.5 for every left/right-wing emoji

words	word_score	Phrase_score	Emoji_score	Overall_score	Political_leanings
[wwgwga, blood, bought, god, first, wife, amer...	22	0.0	0.0	22.0	Right Wing

Figure 2.4: Second scoring system

When using the same example as before the score is now +22

Third scoring system: - Strong and weak terms are given the same weight.
-/+ 1 for every strong left/right-wing term
-/+ 1 for every weak left/right-wing term

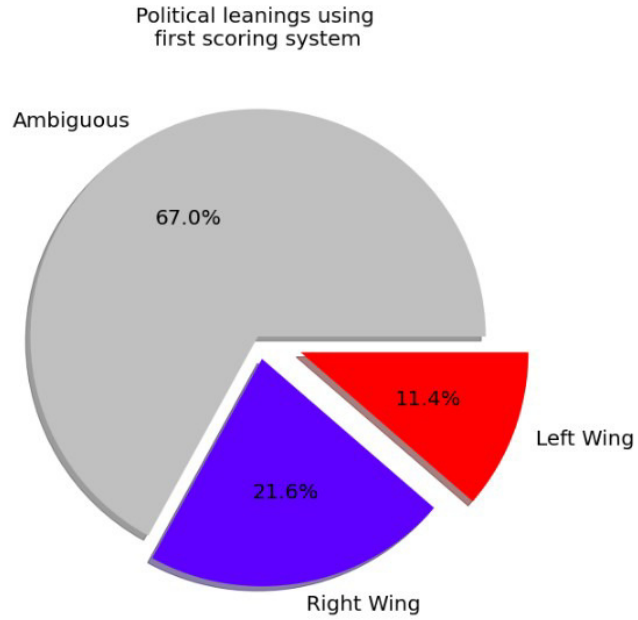
-/+ 1 for every strong left/right-wing phrase
-/+ 1 for every weak left/right-wing phrase

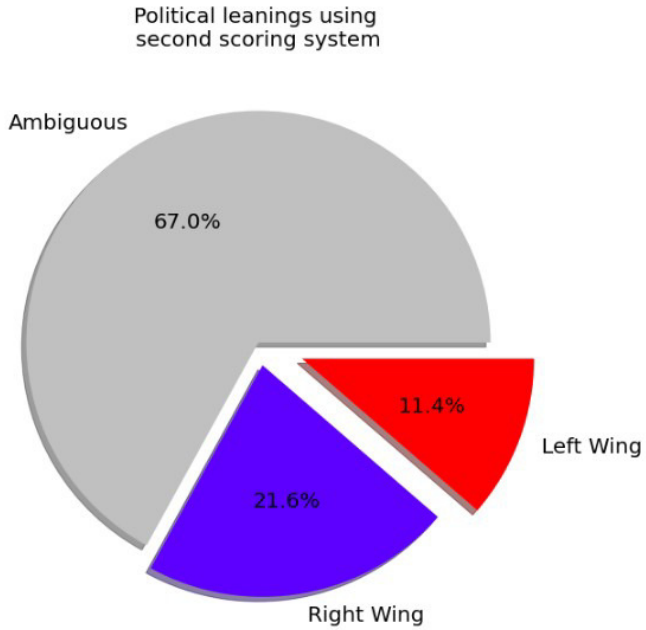
-/+ 0.5 for every left/right emoji

When using the same example as before the score is now +4.

words	word_score	Phrase_score	Emoji_score	Overall_score	Political_leanings
[wwgwga, blood, bought, god, first, wife, amer...	4	0.0	0.0	4.0	Right Wing

Figure 2.5: Third scoring system





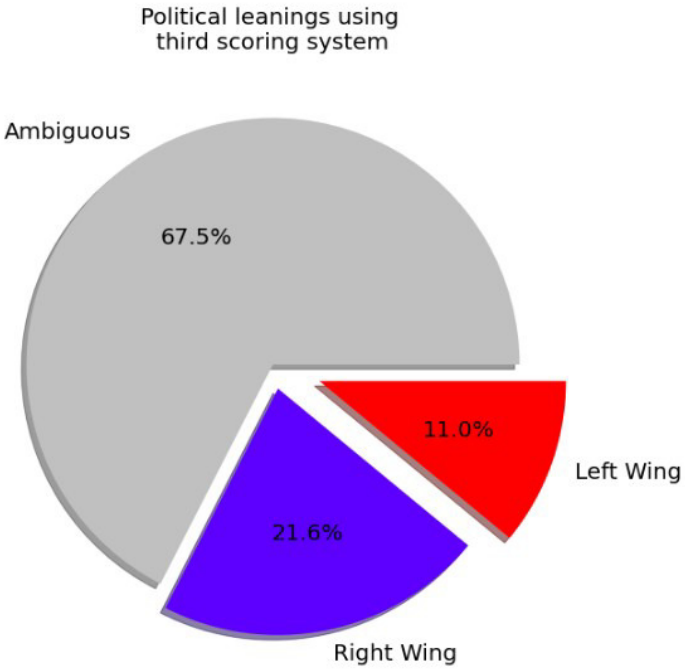


Figure 2.6: Results of different scoring systems

Looking at the results of the three different scoring systems in figure 2.6 it is clear there is very little variation in the results. In fact, there was no difference in the results between the first and second system. Meanwhile, there was only a very slight change in the third system, where the right-wing group remained virtually unchanged and the left-wing group shrank by 0.4%. This suggests then that the numbers used in the scoring are not as important as first assumed.

Verifying the algorithm

In order to actually prove the algorithm is working correctly, it would be wise to analyse what the left and right-wing accounts are retweeting. This should give an indicator as one would expect the right-wing accounts to be retweeting politicians such as Donald Trump and other conservative figures, whilst one would expect the left-wing set to be retweeting more liberal figures and social activists.

The tests will be conducted on days in the July dataset, where there tends to be around 800,000 tweets for each day. Additionally, the test will also be run on data collected recently i.e. data which has not been provided by CSRI. The algorithm will be using the first scoring system.

The top 10 retweets from both the left and right-wing accounts will be researched and given a brief summary, i.e. their background (if possible), the content of their tweets, their twitter description etc. This should give an indication as to whether an account is left or right-wing.

13th July 2020- right-wing set

	account_retweeted	counts
0	@RealJamesWoods	20422
1	@ACTBrigitte	7574
2	@realDonaldTrump	4292
3	@SheepKnowMore	4068
4	@gatewaypundit	2637
5	@cjtruth	1602
6	@catturd2	1547
7	@SweetSoaps	1505
8	@WarNuse	1409
9	@orbitxblink	1391

- @RealJamesWoods - the account of the actor James Woods, is a registered Republican and a staunch Donald Trump supporter (Folley 2020).

- @ACTBrigitte – the account belongs to Brigitte Gabriel, a conservative author who is also a strong supporter of Trump. She is also the founder of the anti-Muslim group “ACT! for America” (Beinart 2018).

- @realDonaldTrump - this is this the account belonging to the President of the USA, Donald Trump.
- @SheepKnowMore – this account has now been suspended, but there are screenshots of this account expressing support for “Obamagate” located online. “Obamagate” is an unfounded conspiracy theory which accuses the Obama administration of a cover-up, relating to investigation into collusion between Trump’s election campaign and Russia. This theory has been promoted by Trump himself as well as other right-wing figures.
- @gatewaypundit - the twitter account for the far-right website, The Gateway Pundit, which has promoted debunked conspiracy theories amongst other misinformation (Tani 2017).
- @cjtruth - a right-wing conspiracy theorist and believer of the “Deep State”. Describe themselves as a “*Patriot/Digital Soldier*”.
- @SweetSoaps – the user’s description reads, “*#God #MAGA #AMERICA #patriot #Trump2020… #savethechlidren*”. #savethechildren refers to the QAnon conspiracy. This user is therefore promoting the baseless rumour that the world’s elite are running a child sex trafficking ring (The Sun 2020).
- @WarNuse - another right-wing conspiracy account, who again, has promoted QAnon. Their account has now been suspended.
- @orbitxblink - this account had over 1391 retweets from right-wing accounts; however, they appear to be left-wing. The content of the tweet that was being retweeted reads - “*capitalist propaganda has rotted ur brains*”, suggesting the user is a socialist. This likely occurred due to the small margin of error involved in the algorithm. Figure 2.7 displays the number of left and right-wing users who have retweeted @orbitxblink. The number of retweets from the left-wing accounts is nearly 14 times higher than that of the retweets from the right-wing accounts. This is clearly a significant difference and one which we would expect from a tweet criticising capitalism.

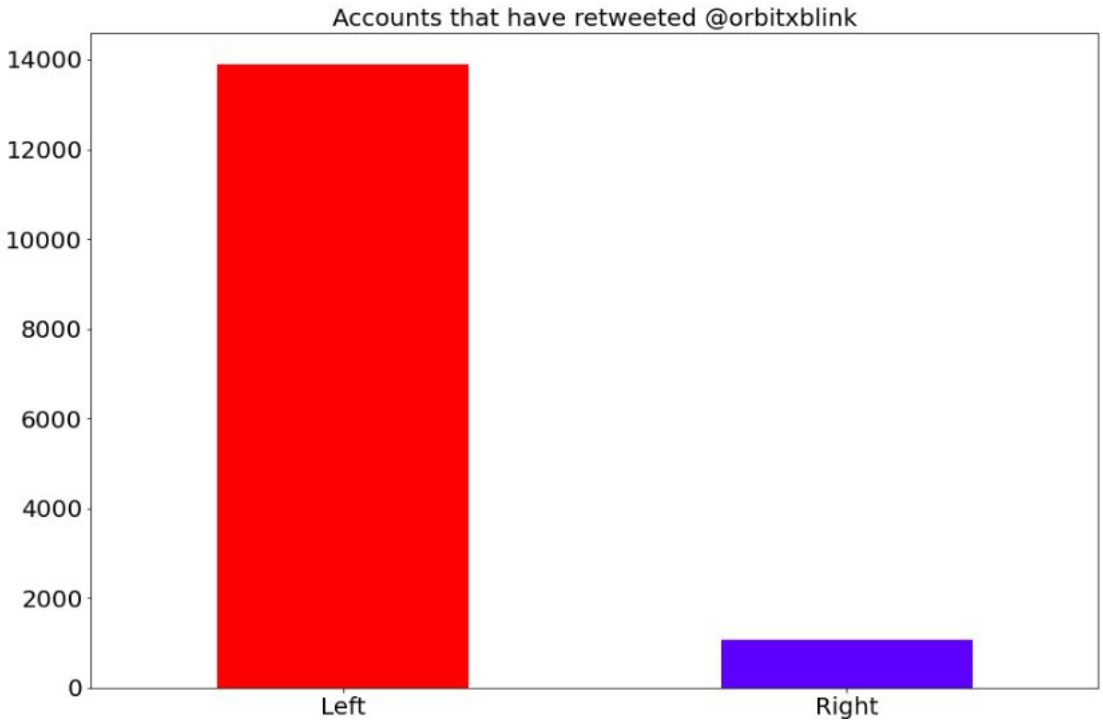


Figure 2.7: Accounts that have retweeted @orbitxblink

13h July 2020- left-wing set

	account_retweeted	counts
0	@orbitxblink	13589
1	@Y00NSTH0T	3229
2	@kylegriffin1	1795
3	@tribelaw	1525
4	@JRehling	1408
5	@DeanObeidallah	1024
6	@skaijackson	862
7	@confusedophan	831
8	@GeorgeTakei	720
9	@RealJamesWoods	704

- @orbitxblink - since the account posted a tweet that was critical of capitalism it is not surprising many left-wing accounts retweeted this.
- @YOONSTH0T - the account has been suspended. The tweet from the account reads: *"normalize deleting tweets that spread misinformation instead of keeping it up for clout."* Unfortunately this does not give a clear indication that they are left-wing.
- @kylegriffen1 - is a journalist from the US cable channel MSNBC. He has constantly criticised Trump and his policies on his account.
- @tribelaw - the legal scholar, Laurence H Tribe has just 1500 retweets from left-wingers. Tribe has a history or criticising Trump and was vocal in his support of Trump's impeachment (Tribe 2019).

- @GeorgeTakei – the account for George Takei, the Star Trek actor. Takei has been a social activist for much of his career and has spoken out against President Trump's rhetoric regarding immigration (Chang, Riegle and Effron 2019).
- @RealJamesWoods – has the tenth highest retweets in the left-wing set for the 13th July. As already explained, this account had the highest amount of retweets amongst the right-wing set, and for good reason, as Woods is a registered Republican and Trump supporter.

So why is James Wood receiving retweets from left-wing accounts? Figure 2.8 shows the tweet by James Woods in question. The tweet is clearly expressing his disdain for mainstream media.

When Woods is referring to "mainstream media", he most likely referring to "liberal bias" many conservatives believe mainstream media possesses (Flood, 2020). It could be argued that the left-wing accounts who retweeted did not realise this; however, it is more likely that it is simply the margin of error in the algorithm. Figure 2.9 (Woods 2020) shows the difference between the number of retweets on Wood's account between the left and right set, with him receiving more than 20 times more retweets from right-wing accounts.

Figure 2.8: James Woods' tweet

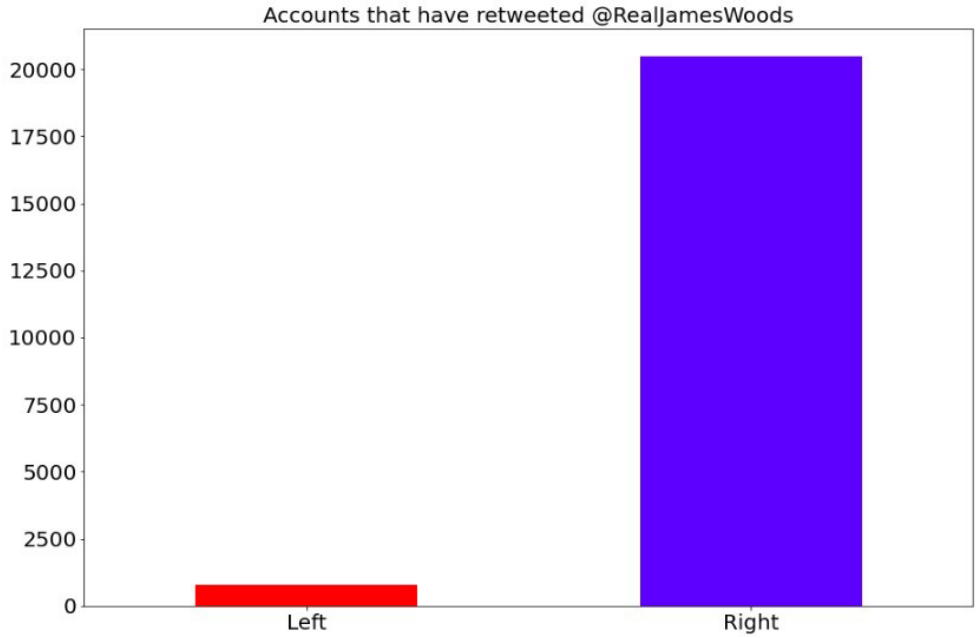


Figure 2.9: Accounts that have retweeted @RealJamesWoods

16th July – right-wing set

	account_retweeted	counts
0	@TheBabylonBee	14010
1	@SheepKnowMore	4999
2	@newtgingrich	4243
3	@cjtruth	3356
4	@WarNuse	3296
5	@JamesOKeefeIII	2920
6	@catturd2	2900
7	@realDonaldTrump	2874
8	@danhill2011	1300
9	@thebradfordfile	965

- @TheBabylonBee – the account for The Babylon Bee, which describes itself as "Christian News satire" (The Babylon Bee 2020) The conservative cable channel Fox news, praised them for "lampooning Democratic politicians and liberal media outlets" (Wulfsohn 2020).
- @SheepKnowMore - the suspended QAnon supporting account, appears again with the second most amount of retweets.
- @newtgingrich – the account for Newt Gingrich, a former speaker of the House of Representatives. Gingrich is a Republican party member and holds many conservative positions. He has been accused of creating the highly partisan political environment that we see in America today (Coppins 2018).
- @WarNuse – suspended right-wing conspiracy theorist.

- @cjtruth - a right-wing conspiracy theorist and believer of the “Deep State”. Describes themselves as a “*Patriot/Digital Soldier*”.
- @JamesOKeefelll - the account for conservative political activist James O’Keefe. He is described by *The Atlantic* (Gray and Coppins 2017) as “once a right-wing media darling”.
- @catturd2- the account appears to be a Trump-supporting troll account. Figure 3.0 (Catturd 2020) shows the user mocking Democrat Presidential nominee Joe Biden, whilst also retweeting Donald Trump.

Figure 3.0: Catturd tweets

- @realDonaldTrump - current US president.
- @danielhill2011 – the user of this account is a Trump supporter, with this Twitter description reading *“Support Pres. Trump! MAGA! KAG! PROUD TO RESIDE IN THE BASKET OF DEPLORABLES”*. “Basket of deplorables” refers to a speech a given by Hillary Clinton, referring to half of Trump supporters as a “basket of deplorables”, a term which would be reappropriated by Trump and his election campaign (CNN 2016).
- @thebradfordfile – this account is a supporter of Trump, who frequently tweets of support of him, as shown in figure 3.1.

Figure 3.1: thebradfordfile tweets

16th July – left-wing set

	account_retweeted	counts
0	@PRIMAGIRIS	21806
1	@ItsDanaWhite	2611
2	@TheBabylonBee	1189
3	@swilkinsonbc	1161
4	@ASlavitt	818
5	@BeckettUnite	781
6	@MalanasQueendom	671
7	@HackneyAbbott	387
8	@Yair_Rosenberg	333
9	@davebancroft	306

- @PRIMAGIRIS – is by far the most retweeted account by left-wingers. However, the account does not seem to express any political opinion, and the tweet that has been retweeted tens thousands of times relates to showing alternative YouTube channels to the YouTuber Shane Dawson. Despite this, for the account's location, the user has put "they/he/she", which could indicate them being left-wing as these are a list of pronouns which is more often associated with the left (Berg 2019).
- @ItsDanaWhite - the user on this account has listed the pronouns they/them in their description, therefore suggesting they are left-wing. In addition, the tweet here expressed support for universal housing to combat homelessness, making it likely that they are socialist.
- @TheBabylonBee received over 1,000 retweets from left-wing accounts, but as already explained they are a conservative satire website. At first this is unexpected, but the results need to be put into context. In the data, there were over 30,000 retweets of this account and less than 4% of these were from left leaning accounts. Figure 3.2 shows that the right-wing retweeted @TheBabylonBee around 14 times more than the left did.

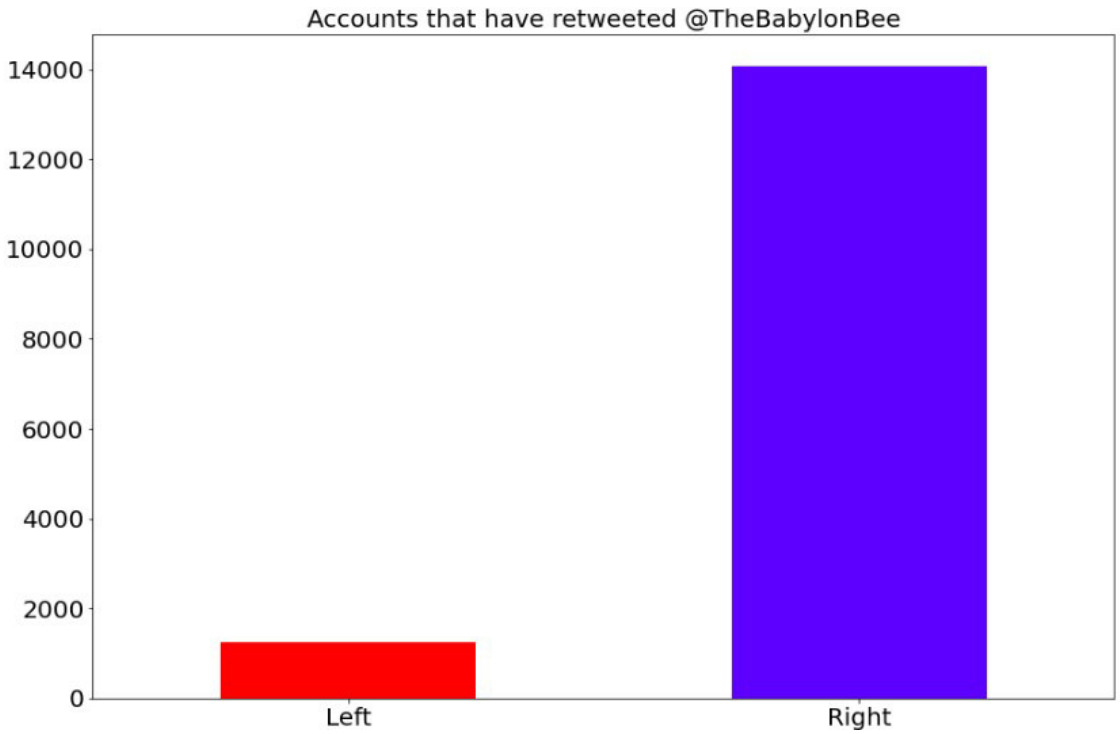
- 

Figure 3.2: Accounts that have retweeted @TheBabylonBee

- @swilkinsonbc –in the dataset, the user of this account tweeted *"Palestine is no longer a place according to google — ethnic cleansing, murder, land theft & corporate conspiracy, all done in israel's name"*. The account user is a campaigner for Palestine, who has gained significant prominence on the left in recent times (Eichler 2015).
- @ASlavitt - the account for Andy Slavitt, the former Acting Administrator of the Centers of Medicare and Medicaid, as well as the head of the Affordable Care Act (ACA) under President Barack Obama. Slavitt has been highly critical of Trump and his handling of the Covid-19 pandemic (Slavitt, 2020), which explains why left leaning accounts have been retweeting him.
- @BeckettUnite – the twitter account of Howard Beckett, described by the BBC (2020) as the left-wing assistant general secretary of Unite (Labour's largest affiliated trade union.
- @MalanasQueendom is certainly left-wing, with their description reading *"Black, queer, & disabled til the rainbow burns the stars out in the sky. Anti-zionist Jew of color. Awkward except when I pop and lock. she/her"*.
- @HackneyAbbott is the Labour MP Diane Abbott, a close ally of former Labour leader Jeremy Corbyn.

- The journalist, Yair Rosenberg, who has 333 retweets from left leaning accounts, has posted in support Joe Biden, and has been very critical of Trump (Rosenberg 2020).
- @davebancroft – the account tweeted *"Do we still have to 'Respect the Referendum result' if the vote was subject to interference? Asking for more than 16 million UK voters..."*. This account is clearly against Brexit, but this alone does not prove they are left-wing, as there are those on the right who are pro-E.U. as well. However, on their account they have consistently criticised the British Conservative Party.

20th July – right-wing set

	account_retweeted	counts
0	@ACTBrigitte	5173
1	@kirstiealley	5095
2	@Lrihendry	3852
3	@MajorPatriot	3302
4	@FOOL_NELSON	2312
5	@martingeddes	2187
6	@WarNuse	2149
7	@Jordan_Sather_	2101
8	@SidneyPowell1	1939
9	@DiamondandSilk	1900

- @ACTBrigitte – account belonging to Brigitte, who as already mentioned, is a conservative author and strong Trump suppoter.
- @kirstiealley- the account for the actress Kirstie Alley. Alley endorsed Trump for the 2016 Presidential Election (Trudo 2016).
- @Lrihendry – an account which has just under 4000 retweets by right-wing accounts in the dataset. The account belongs to Lori Hendry, described by the Daily Mail (2020) as a "digital keyboard warrior" who's Twitter description reads "*Here for @readDonaldTrump•Proudly RTed by President Trump•purveyor of truth... "*.
- @MajorPatriot – a strong supporter of Trump, and regularly retweets right-wing conspiracies.
- @FOOL_NELSON – this account tweeted consistently retweets criticisms of Joe Biden (FOOL_NELSON 2020), whilst retweeting support for Trump, see figure 3.3.

Figure 3.3: FOOL_NELSON retweet

- @martingeddes – belongs to Martin Geddes who appears to be a QAnon believer, and has even written about the subject.
- @WarNuse – suspended conspiracy theorist who has tweeted in support of QAnon.

- @Jordan_Sather – A QAnon believing conspiracy theorist, figure 3.4 shows him tweeting this (Jordan_Sather 2020).

- @SidneyPowell1- the account for the lawyer Sidney Powell, who defended Michael Flynn, former national security advisor in the US, after he was accused of lying to the FBI. She has received praise from Trump, and even had private conversations with him (NY Times 2020).

Figure 3.4: @Jordan_Sather_ tweet

- @DiamondandSilk – the account for two American bloggers and political activists, who have in their Twitter description state, "*President Donald J Trump's Most Loyal Supporters*". They were also Fox Nations hosts, before being dropped for spreading misinformation regarding the coronavirus pandemic (Cillizza 2020).

<u>20th July- left-wing set</u>

	account_retweeted	counts
0	@simimoonlight	7474
1	@RepAdamSchiff	2338
2	@donwinslow	1096
3	@GrimKim	1082
4	@funder	1055
5	@Sifill_LDF	895
6	@ddale8	782
7	@ThePubliusUSA	741
8	@BreeNewsome	692
9	@FrankFigliuzzi1	644

- @simimoonlight – the account tweeted, "*Kanye being bipolar doesn't excuse his spread of misinformation but it does mean how we engage and speak about him shouldn't be ableist.*" The tweet is in reference to the rapper Kanye West's comments regarding the abolitionist Harriet

Tubman, who was born into slavery and went on to free approximately 70 enslaved people during the 1800s. West stated that she “never actually freed the slaves” (Reddick 2020). Tubman is viewed as a progressive hero who fought racial injustice and the oppression of women (Rans 2008). Due to the fact that this account is calling out West for spreading misinformation regarding Tubman the account is most likely left-wing, as fighting social injustices is an issue more associated with the left.

- @RepAdamSchiff is the account for Adam Schiff a member of the Democratic Party, and the US Representative for California’s 28th congressional district. He was chosen to lead the prosecution in Trump’s impeachment trial (BBC 2020), which would explain why those on the left are retweeting him.
- @donwinslow is the account for Don Winslow, the author. He has been a vocal critic of Trump, referring to him as “America’s greatest mistake” (Daily Kos 2020).
- @GrimKim appears to be an anarchist and anti-capitalist, who has tweeted her anger at perceived “union-busting”, see figure 3.5 (Grim Kelly 2020). This suggests that they are a sociallist.

Figure 3.5: GrimKim tweet

- @funder is the account of Scott Dworkin, who according to his website (2020) is a “proud member of #theResistance”, the group who oppose the Presidency of Trump.
- @Sifill_LDF is the account for Sherrilyn Ifill a law professor and president and director-counsel of the NAACP Legal Defense Fund (NAACP 2020). Being part of a progressive organisation one would expect left leaning accounts to be retweeting her.

- @ddale8 is a CNN journalist who fact checks Trump on his Twitter account. According to Dale, Trump has said over 5000 false things as president (Dale 2019).
- @ThePubliusUSA is a frequent critic of Trump, referring to him as a "traitor". Additionally their description states they are a believer in "equality", which hints at left-wing ideals.
- @BreeNewsome belongs to Bree Newsome, who in 2015 climbed up a flagpole on the South Carolina state house and removed the Confederate flag flying there. This act made her a social activist and a hero online to some (BBC 2015).
- @FrankFigliuzzi, the account for Frank Figliuzzi, the former Assistant Director of the FBI's Counter-intelligence Division (FBI 2011). Figliuzzi described Trump as "the Greatest threat to the country" (MSNBC 2020). It is no surprise that left leaning accounts are retweeting him.

Analysis of the results from the data

The results do seem to follow a pattern. The most frequently retweeted accounts by right-wingers are made up of strong Trump supporters, sometimes Trump himself, conspiracy theorists and Republican politicians. Meanwhile, the most frequently retweeted accounts by left-wingers are social activists, accounts who criticise Trump and his policies and Democrat politicians. It should be stressed that not everyone is in the left-wing set of most retweeted accounts are left-wing, especially those in the Democratic Party. However, they are further left than than Trump and the Republican Party.

13. Verifying the algorithm on data collected using tweepy

The algorithm will now be tested on data collected which does not have a set of search terms focussed upon misinformation. Using the tweepy Python library over 100,000 recent tweets were collected. The same left or right algorithm was used on this new dataset that was applied to the other data.

Top 10 accounts retweeted by the right-wing accounts

	account_retweeted	counts
0	@realDonaldTrump	174
1	@dbongino	89
2	@charliekirk11	43
3	@catturd2	40
4	@EricTrump	40
5	@RealJamesWoods	40
6	@Jkruge	36
7	@JackPosobiec	26
8	@marklevinshow	21
9	@stillgray	19

- @realDonaldTrump – President Trump has the most retweets from right-wing accounts, which is hardly surprising given what has been discussed.
- @dbongino - an account belonging to Dan Bongino, an American political commentator. Bongino unsuccessfully ran for Congress as a Republican in Florida 19th District in 2016 (Scott E 2016). He is a strong supporter of Trump (Schwartz 2018), and in an interview in 2018, he stated "My entire life right now is about owning the libs. That's it" (Amatulli 2018).
- @charliekirk11 - the account belonging to Charlie Kirk, the founder and president of Turning Point USA, an American conservative non-profit organisation. Those who back him believe he is the future of conservative politics (Nelson 2015).
- @catturd2 – this account also appeared in the 16th July dataset, having over 2000 retweets from right-wingers. As mentioned there, the account appears to be a Trump supporting troll account.
- @EricTrump – Eric Trump, the son of President Donald Trump. He helped promote his father's presidential campaign in 2016 (Revesz 2016).
- @RealJamesWoods – The actor James Woods was the most frequently retweeted account by right-wingers in the 13th July dataset. As stated previously, he is a staunch supporter of Donald Trump.
- @Jkruge – The account describes themselves as a "TRUMP WARRIOR".
- @JackPosobiec – this account belongs to Jack Posobiec, a political activist belonging to the alt-right. He has worked with white supremacists, neo-fascists and antisemites (Hayden 2020).
- @marklevinshow – this account belongs to Mark Levin, the lawyer and author. He worked under the Reagan administration and has been been described as being right-wing by CNN (Stelter 2017).
- @stillgray – this account belongs to Ian Miles Cheong. The tweet that was retweeted by 19 right-wing accounts read "*This female LASD deputy is a hero. She's helping her partner*

tend to his wounds while she's bleeding from the jaw. She was shot in the face. Blue lives matter." The group Blue Lives Matter was started in response to the group Black Lives Matter. Blue Lives Matter advocates those who kill law enforcement officers should be tried of a hate crime (Business Insider 2017). The Blue Lives Matter group has been supported mainly by those on the right, with Trump and other conservatives embracing them in the lead up to 2016 Presidential election (NBC 2016).

Top 10 accounts retweeted by the left-wing accounts

	account_retweeted	counts
0	@ilyseh	43
1	@MeidasTouch	32
2	@kylegriffin1	22
3	@chartdata	18
4	@pant_leg	16
5	@jamzenn	16
6	@ChrisEvans	16
7	@NYGovCuomo	16
8	@Marisa_Ingemi	15
9	@ceetener	14

- @ilyseh – is the account for Ilyse Hogue, the president of NARAL Pro-Choice America. The group opposes restrictions on abortions.
- @MediasTouch- An American political action committee with the purpose of stopping the re-election of Donald Trump (Moran 2020).
- @kylegriffen1 – This account was the third most retweeted account from left-wingers in the 13th July dataset. He is a journalist from the US cable channel MSNBC. He has constantly criticised Trump and his policies on his account.

- @chartdata – The account for Chart Data, a company which promotes updates on music artists, tweeted "'*MAP OF THE SOUL: 7' remains at #1 on the World Albums chart for a 22nd week.*" This tweet has no political messages or connotations, so will be ignored.
- @pant_leg – This account tweeted the picture in figure 3.6 (pant_leg 2020). This suggests the user is a feminist and most likely left-wing. They also have the pronouns "she/her" in their description, which, as previously discussed, is a trait associated with some left-wingers.

Figure 3.6: pant_leg *tweet*

- @jamzenn – the user of this account is an artist. Although they have not tweeted anything relating to politics, their description contains the pronouns "*They/them*", suggesting they are left-wing.
- @ChrisEvans – the account of the actor Chris Evans, known for his portrayal of Captain America in the Marvel films. Evans tweeted encouraging people to vote in 2020 Presidential Election. Evans has expressed some progressive values such as supporting gay marriage (The Huffington Post 2016). He has also been critical of Donald Trump (Oldham 2020).
- @NYGovCuomo – the account for Andrew Cuomo, who is currently serving as governor of New York. He is a member of the Democratic Party.
- @Marisa_Ingemi – the account is a sportswriter. Their description contains the pronouns "*she/her"*, indicating they are left-wing.
- @ceetenar – this account Twitter name is "*cedar|BLM*", expressing their support for the group Black Lives Matter. They also have the pronouns "*he/him"* in their description. This, again indicates that the user is left-wing.

Analysis of the results from the collected data

The most retweeted accounts by right-wingers in the collected dataset follow a similar pattern as before. The list is made up of Trump supporters, Republicans, and even a right-wing extremist. The accounts most frequently retweeted by the left are, again, made up of those criticise Trump and policies and those who hold progressive beliefs (e.g. pro-choice).

Verdict

The results from both the data provided by the Institute, and the data collected via tweepy, show a clear divide in the accounts retweeted by the left and right-wing accounts. Right-wing accounts are considerably more likely to retweet accounts that are supporters of Trump, strong conservatives and promoters of conspiracy theorists, while the left are much more likely to retweet those who are vehemently against Trump, liberals, socialists and social activists. These results show that the accounts that are getting labelled left or right-wing, are being labelled correctly the overwhelming majority of the time.

As a result, the left or right algorithm can be accurately applied to the vaccination dataset in order to observe differences between the left and right-wing sets when it comes to the analysis of the data.

Chapter 3- Analysis of the Vaccination Dataset

Objective of the this chapter

Now that the left/right algorithm has been successfully created and validated it can be be applied to the vaccination-related dataset. The objective of this chapter is to conduct a thorough analysis of the vaccination-related data in order to determine what misinformation is leading people to distrust vaccines and whether political leanings have an effect on this.

Analyses will be conducted on the following:

- Frequency of vaccination related tweets
- Hashtag used
- Retweets
- N-grams
- Topic modelling
- Left and right-wing comparisons using with all the above

Datasets used

In this chapter the entirety of the data provided by the CSRI has been used, but since the focus of this project is specifically on misinformation regarding vaccinations, the data will need to be filtered so that only tweets where the subject is vaccines are left in the dataset.

Analytic tools used in the chapter

Below are the analytical tools that were used in this chapter. Please refer to the background section for a more detailed description.

- pandas
- regex
- nltk
- sklearn
- matplotlib
- seaborn

Filtering the data

In order to filter the data a list of terms related to vaccines will need to be created which will capture the largest amount of data possible.

Below is the list of vaccination related terms. The Series.str.contains() function has been used to test if an element from the list “anti_vax_terms”, is contained within a string.

```
anti_vax_terms = [ “vax”, “vaccin”, “pharma”, “booster”,
                                    “immun”, “inoculat”, “inject” ]
```

The reason for using this function is that it avoids having to use many different variations of the same base word in the list anti-vax-terms. Shown below are a sample of vaccine-related words that contain an element in anti_vax_terms:

- **vax** – anti vaxxer, anti-vaxxers, anti vaxxer, anti vaxxers , anti-vax, anti vax
- **vaccin** – vaccine, vaccines, vaccination, vaccinations, vaccinated, vaccinating,vaccinate
- **pharma** – pharma, pharmacy, pharmaceutical
- **immun** – immunise, immunisation, immunity, immune
- **booster** - booster
- **inoculat** – inoculating, inoculation, inoculated
- **inject** – inject, injection, injected, injecting

This is a more efficient way of filtering the data to contain only vaccine-related tweets.

Analysing the number of vaccine-related tweets

Figure 3.7 shows the total amount of vaccine-related tweets across seven different weeks in 2020. The number of tweets has been steadily increasing throughout the year, before a massive increase from 15-21 June to 13-19 July where the number of vaccine-related tweets increased by more than 134% to the figure 36121.

Figure 3.7 also displays the proportion of vaccine-related tweets across the 7 different weeks. The later weeks contain a far bigger proportion, with the week in June and July taking up more than 50% of the dataset.

From the first week in 2020, the number of vaccine-related tweets increased by nearly 10 times the amount at the start of the year. This huge increase is likely to be driven by the projected vaccine for coronavirus, but this will need to be confirmed.

Date	**Number of tweets**
01-07 January	3720
09-15 January	6088
04-10 March	8105
19-25 March	11258
26 March - 01 April	11950
15-21 June	15379
13-19 July	36121

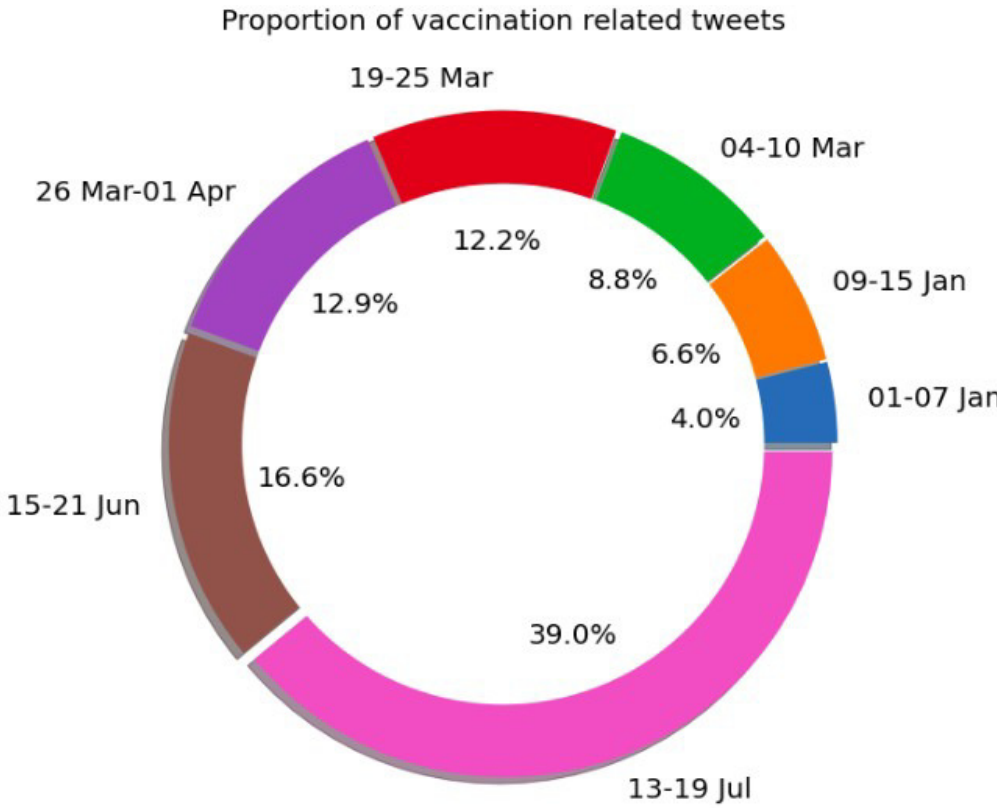


Figure 3.7: Amounts and proportion of vaccination related tweets

In order to confirm that the projected Covid-19 vaccine is the driving force behind the rise in the number of vaccine-related tweets, the Series.str.contains() function will need be used on the vaccine dataset. If the vaccine-related tweet contains any of the elements in the list, corona_search, it will be deemed a Covid-19 vaccine-related tweet. Other vaccines were also tested , such as the flu vaccine, HPV vaccine, MMR vaccine and the DTaP vaccine. However none of these generated significant results.

corona_search = [“covid” , “corona”, “sars”, “chinese virus”*]

*US President Donald Trump referred to the coronavirus as the “Chinese virus” in March 2020 (BBC 2020).

Figure 3.8 display the results. The number of Covid-19 vaccine-related tweets was unsurprisingly non-existent in both of the weeks in January. From March this number rose sharply, increasing by 121% between 04-10 March and 19-25 March. At this point Covid-19 vaccine tweets represented a quarter of all vaccine tweets, which would then rise to 28%, with a total of 3483. These increases fall in line with the total increases, therefore suggesting a positive correlation up to this point. However, in June, the number of Covid-19 vaccine-related tweets declined to 2301, whilst the total vaccine-related tweets continued to increase. In July though, Covid-19 vaccine-related tweets tripled to 7,805, with their proportion rising from 15% to 22%. This significant increase came hand in hand with the large increase in total vaccine-related tweets, which more than doubled from 15,379 to 36,121. It appears that the prospect of a Covid-19 vaccine has led to significant increases in the total number of vaccine-related tweets. Although when Covid-19 vaccine tweets decreased, total vaccine tweets increased, so it is not the only factor.

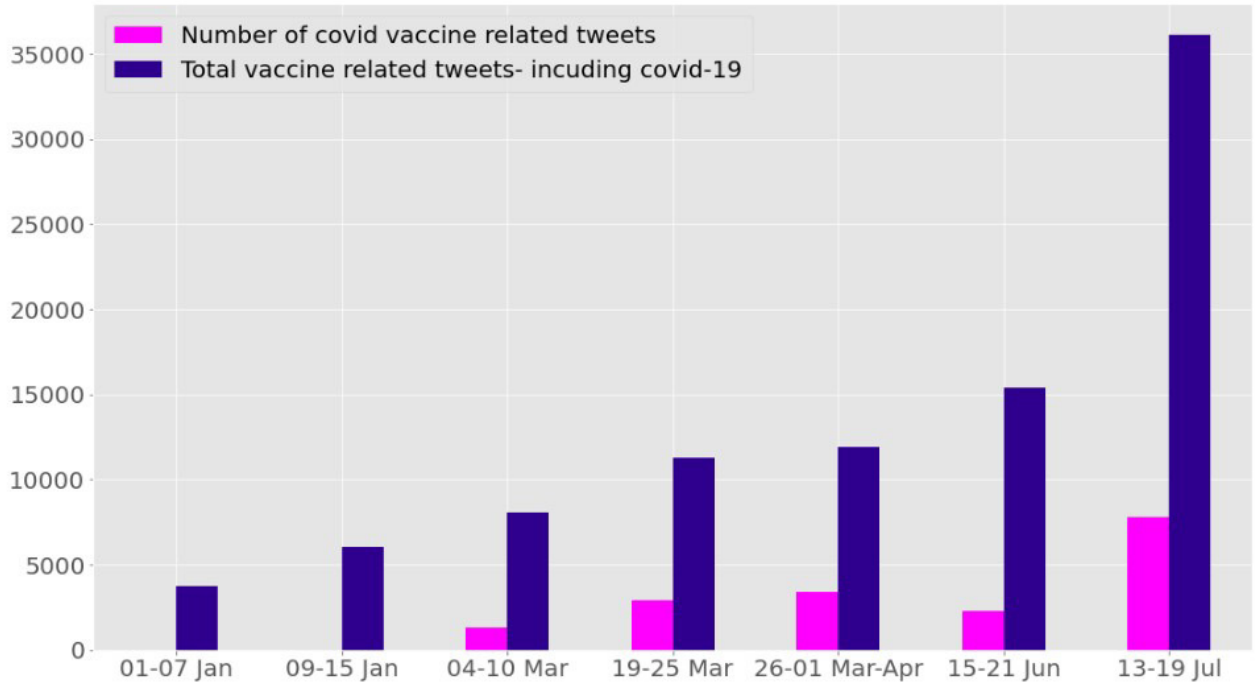


Figure 3.8: Covid-vaccine related tweets

Date	Covid-19-vaccine	Proportion of Covid-19 tweets (%)	Total number of vaccine-related tweets
01-07 January	1	0	3720
09-15 January	5	0	6088
04-10 March	1316	16	8105
19-25 March	2917	26	11258
26 March-01 April	3383	28	11950
15-21 June	2301	15	15379
13-19 July	7805	22	36121

Figure 3.9: Table of results showing proportion of Covid-19 vaccine tweets

Comparing the left and right

The left/right-wing algorithm can be applied to the vaccination dataset. Figure 3.9 displays the ratios of the political leanings in the vaccination dataset. Just over a fifth of the tweets are from right-wing accounts, while under 15% of vaccination related tweets are from left-wing accounts. It would appear that right wing accounts are responsible for more misinformation regarding vaccines. However, a greater analysis is needed.

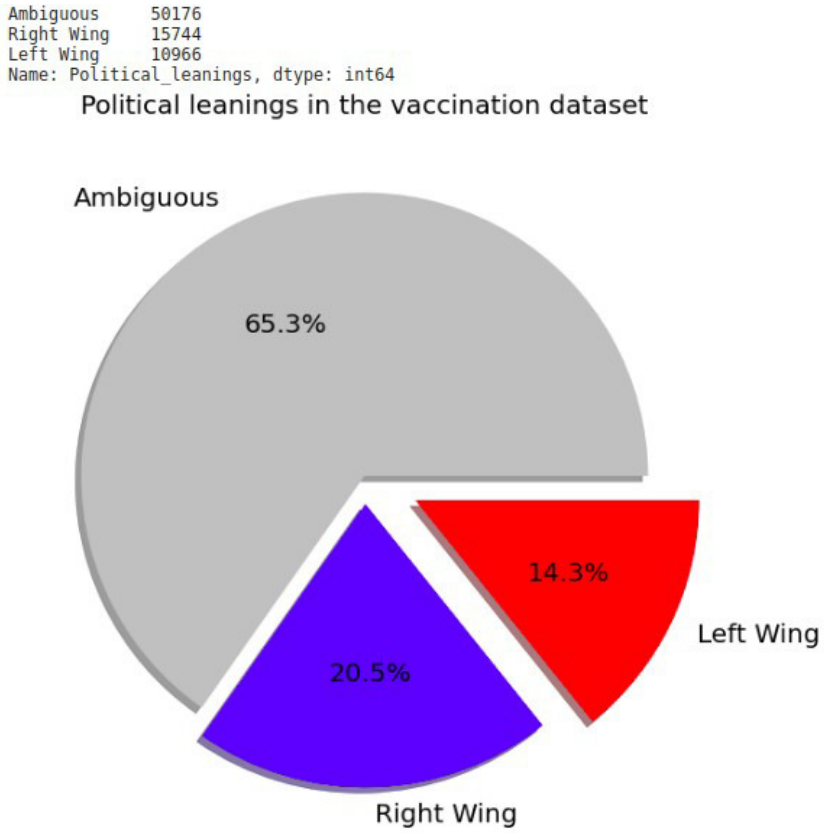


Figure 4.0: Left/right-wing tweets on vaccines

Hashtags

The Twitter Help Centre explains that hashtags are “used to index words or topics” on Twitter. This allows users to follow the topics that they are interested in. If a “hashtagged” word becomes popular enough they can become a trending topic (Twitter 2020). Therefore, by looking at the common hashtags in the data we should be able to obtain a better understanding on what topics are being discussed around vaccinations.

The hashtags in each tweet can be returned using the re.findall(pattern, string) method:

```
return re.findall('(#[A-Za-z]+[A-Za-z0-9-_]+)', text)
```

The code above finds and returns any hashtags in the text, followed by one or more alphabetical characters, then followed by one or more alphanumeric characters. Figure 4.0 shows this in action in the dataframe. If there is no hashtag in the text it will simply return nothing.

text	user_followers_count	Political_Leanings	hashtags
@misssuestar @RWPUSA When the republicans stan...	172.0	Ambiguous	[]
@AlumiLynn @HeadCaseRN @doritmi @silversynergy...	102.0	Ambiguous	[]
RT @PatrioticProgr1: 1) Have you ever called a...	321.0	Ambiguous	[#AntiVaxxer]
RT @PeterHotez: #Flu is taking off now. Still ...	96.0	Ambiguous	[#Flu, #vaccine, #antivax]
@Tappy_95 @TheQuartering look - communism isnt...	42.0	Ambiguous	[]

Figure 4.1: Hashtags column in dataframe

The “hashtags” column is then used to create a new dataframe. This dataframe is then flattened, and then the hashtags that are the same are grouped together with their frequencies. Please note, this is the same process that was performed on the words in the Twitter descriptions in chapter 1.

Figure 4.1 displays the results of the top 10 most frequently used hashtags used in the entire vaccination dataset. The most frequently used hashtags are dominated by conspiracy theories. The most frequently used hashtag is #BigPharma, which refers to a group of conspiracy theories which claim the pharmaceutical industry and large corporations are working in an insidious way, with only an interest in profit. The other claim is they cause or/and exacerbate diseases (Blaskiewicz 2013).

Many of the hashtags here seem to be related to Qanon/Deep state conspiracy theories. As previously mentioned, QAnon is a right wing conspiracy theory which claims the world is being controlled by satanist paedophiles (many of them being Democrats), and President Trump is battling them. Meanwhile the “deep state” refers to a network inside the government which allegedly controls state policy behind the scenes (Dictionary (2020). These conspiracy theories do share some similarities. According to Vox (2020), many of Trump’s supporters believe the coronavirus pandemic is a “deep state coup” which is trying to oust Trump. This would explain the hashtag “#DeepStateVirus”. It should be noted that six of the hashtags in this list have similar frequencies i.e. from #Qarmy to #TheGreat; this suggests that these hashtags are appearing together in the same tweet which has been retweeted multiple times.

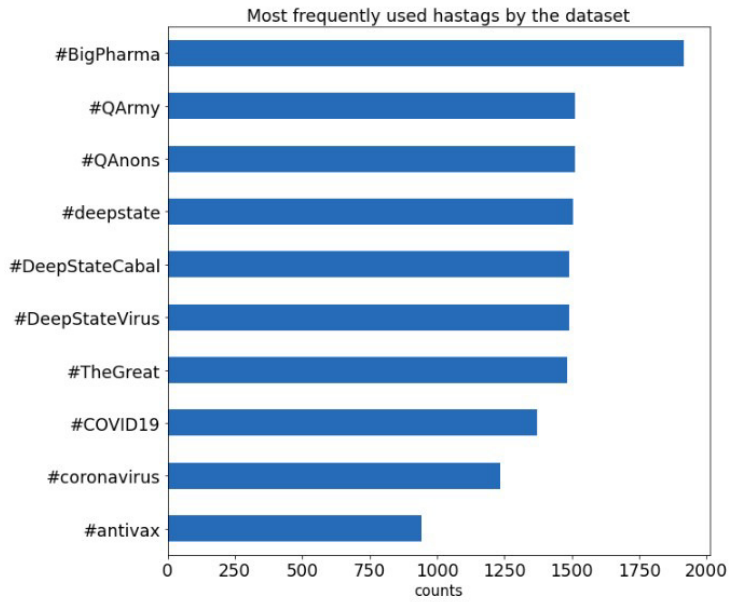


Figure 4.2: Most used hashtags

Figure 4.2 displays the most frequently used hashtags amongst left and right-wingers. The first observation is that right-wing accounts are using hashtags at a higher frequency than left-wing accounts. The second observation is that the list of right-winger hashtags is almost identical to the overall set, again suggesting that the right-wing are producing most of the misinformation regarding vaccines.

The similar frequencies in the overall and right-wing set suggest there are strong correlations between the most frequently used hashtags. To find the extent to which these hashtags are correlated to each other, the hashtag must be turned into vectors, in the identical way the words in the Twitter description in the previous chapter were. For efficiency, only hashtags that appear at least 500 times will be correlated.

The correlation matrix can then be plotted as a heatmap using the seaborn package. The results of this are shown in figure 4.4.

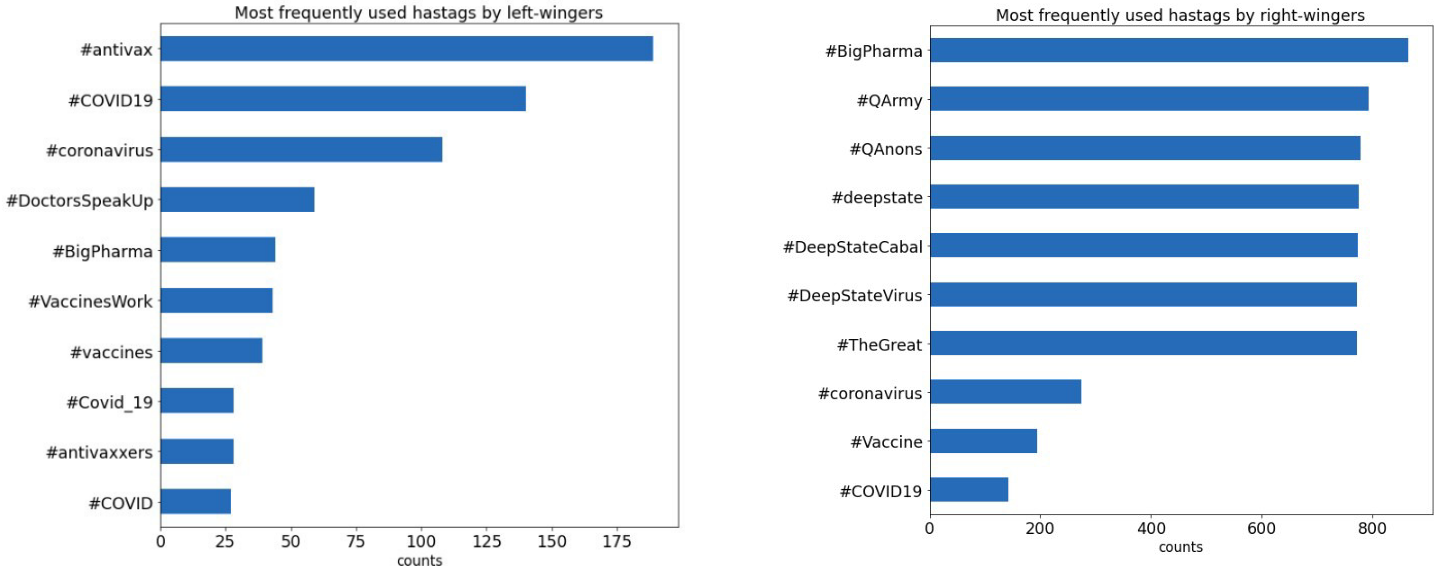


Figure 4.3: Most used hashtags by the left and right

Figure 4.4 displays the hashtags correlations as a heatmap. There are very strong correlations between the top seven (#BigPharma to #TheGreat) hashtags in figure 4.3, with the correlations being almost 1.00, which would signify a perfect correlation. These seven hashtags also possess a negative correlation with other hashtags. Therefore, when one of these seven hashtags appears, they almost all appear together, and not with other hashtags. After analysing which accounts are tweeting these hashtags it was found that a tweet from the, now suspended, account @shoptaraeveland had been retweeted 1488 times, and the tweet contained all seven of the most frequent hashtags.

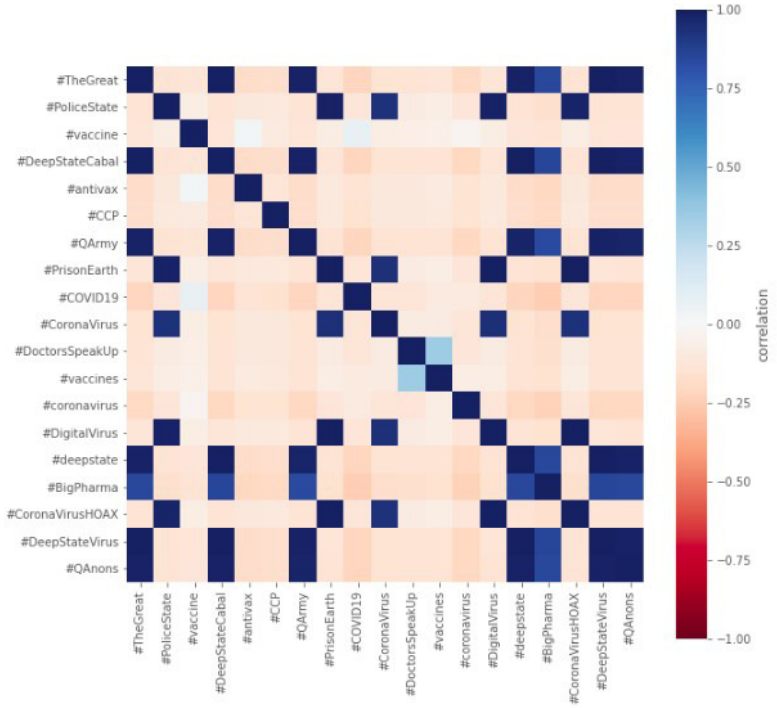


Figure 4.4: Heatmap of hashtags

Retweets

As discussed in the introduction, one of the reasons that misinformation spreads so quickly is due to retweeting. A retweet is simply a reposting of a tweet that allows a user to quickly share a tweet with all their followers (Twitter, no date given). It is a very simple feature to use, so it is easy to see why a post can spread so rapidly through it.

In the data, a retweet will have “RT” to signify that it is a retweet. In order to identify a retweet a new column is created using the following code:

```
Total_Vax_twitter_political['is_retweet'] =  Total_Vax_twitter['text'].apply(lambda x: x[:2]=='RT')
```

The code above checks the first two characters in a tweet, if they are “RT”, then the tweet is a retweet and the column will be set to “True”. The table in figure 4.5 shows the results in the vaccination dataset. Just under half of all tweets in the data are retweets.

	Counts	Percentage %
Retweet	39834:	44
Not a retweet	50790	56

Figure 4.5: Table of retweets proportions

Figure 4.6 shows the number and proportion of retweets amongst the left and right-wing accounts. More than half, 57%, of tweets from right-wing accounts are retweets whilst only 39% of tweets from left-wing accounts are retweets. This suggests that right-wing accounts are spreading misinformation more than left-wing accounts. This is also supported by the findings of the most frequent hashtags, as the overwhelming majority of occurrences of the most frequent hashtags were from retweets of one account.

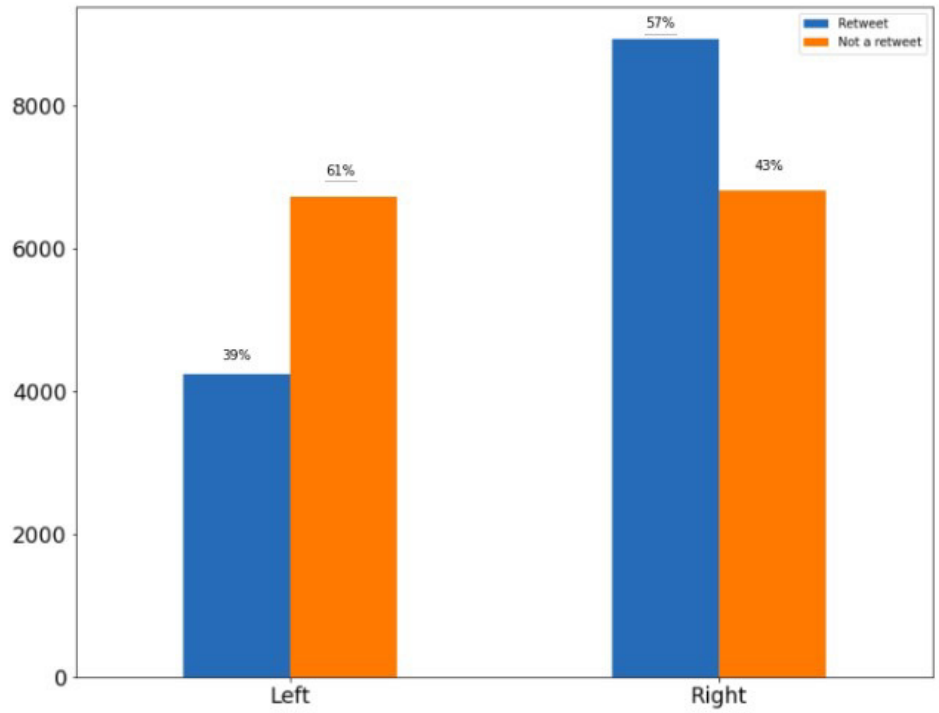


Figure 4.6: Left and right retweets

The relationship between followers and retweets

This section will be analysing the top 15 most retweeted accounts in the vaccination dataset, and whether there is a link between retweets and number of followers on an account. In order to return which account is being retweeted the following code was used:

```
return re.findall('(?<=RT\s)(@[A-Za-z]+[A-Za-z0-9-_]+)', text)
```

The lookbehind (?<=RT\s) asserts that at the current position in the string, what precedes is the characters “RT” and one white-space. If the assertion succeeds, the expression will match the retweeted account name.

The table in figure 4.7 displays the number of retweets, number of followers and the retweet:follower ratio for the top 15 most retweeted accounts. Please note, the tweets that were being retweeted from the accounts @CNN, @roccogawlatilaw and @rmayemsinger were not in the dataset, therefore their follower counts had to be added in manually.

	account_retweeted	number_of_retweets	follower_count	Ratio
0	@cjtruth	2419	169971	0.014232
1	@Jordan_Sather_	1892	177613	0.010652
2	@shoptaraeveland	1479	1275	1.160000
3	@ChildrensHD	848	35760	0.023714
4	@PeterHotez	799	74487	0.010727
5	@va_shiva	685	127703	0.005364
6	@CNN	522	49075837	0.000011
7	@EwdatsGROSS	491	130392	0.003766
8	@roccogalatilaw	455	17200	0.026453
9	@TomFitton	452	1087297	0.000416
10	@rmayemsinger	421	107400	0.003920
11	@LotusOak2	416	18240	0.022807
12	@RichHiggins_DC	403	30704	0.013125
13	@dockaurG	389	13862	0.028062
14	@HoodHealer	369	34223	0.010782

Figure 4.7: Follow:retweet ratio

The account @cjtruth has the most amount of retweets. This account came up frequently when verifying the left and right-wing algorithm, as did the account Jordan_Sather_. Both of these accounts have posted numerous conspiracies regarding vaccinations. Figure 4.8 displays the data from the table in a scatter-graph allowing for a better visualisation of the results.

The graph shows no real correlation between number of followers and number of retweets. In fact, the black dot in the top left represents CNN, who have nearly 50 million followers but only had 500 around retweets in the data. As a result of this, their retweet:follower ratio is extremely low hence their dot being coloured as black. Interestingly, CNN's tweet was critical of the anti-vaccination movement, which could be a reason why it had so few retweets when considering CNN's massive following.

The other account with a low ratio is @TomFitton, who possesses over a million followers but had less than 500 retweets. Meanwhile, the account with the third highest amount of retweets was @shoptaraeveland, which actually had the lowest amount of followers in the list, therefore giving it the highest ratio. It was the only account to have a ratio greater than 1.

With by far the highest ratio (the yellow dot in figure 4.8), the account @shoptaraeveland is very much against vaccinations. The user posted a link in one of their tweets which redirects to Del Bigtree's Facebook page. Bigtree is an America producer and CEO of the anti-vaccination group "Informed Consent". He produced a documentary which featured the discredited former physician, Andrew Wakefield (Pimlott 2019). Despite Bigtree's vaccination conspiracy theories being debunked, he still draws some public support, which is highlighted by the fact that people are retweeting a link to his content. This user also seems to believe that the coronavirus pandemic has been created by the "deep state".

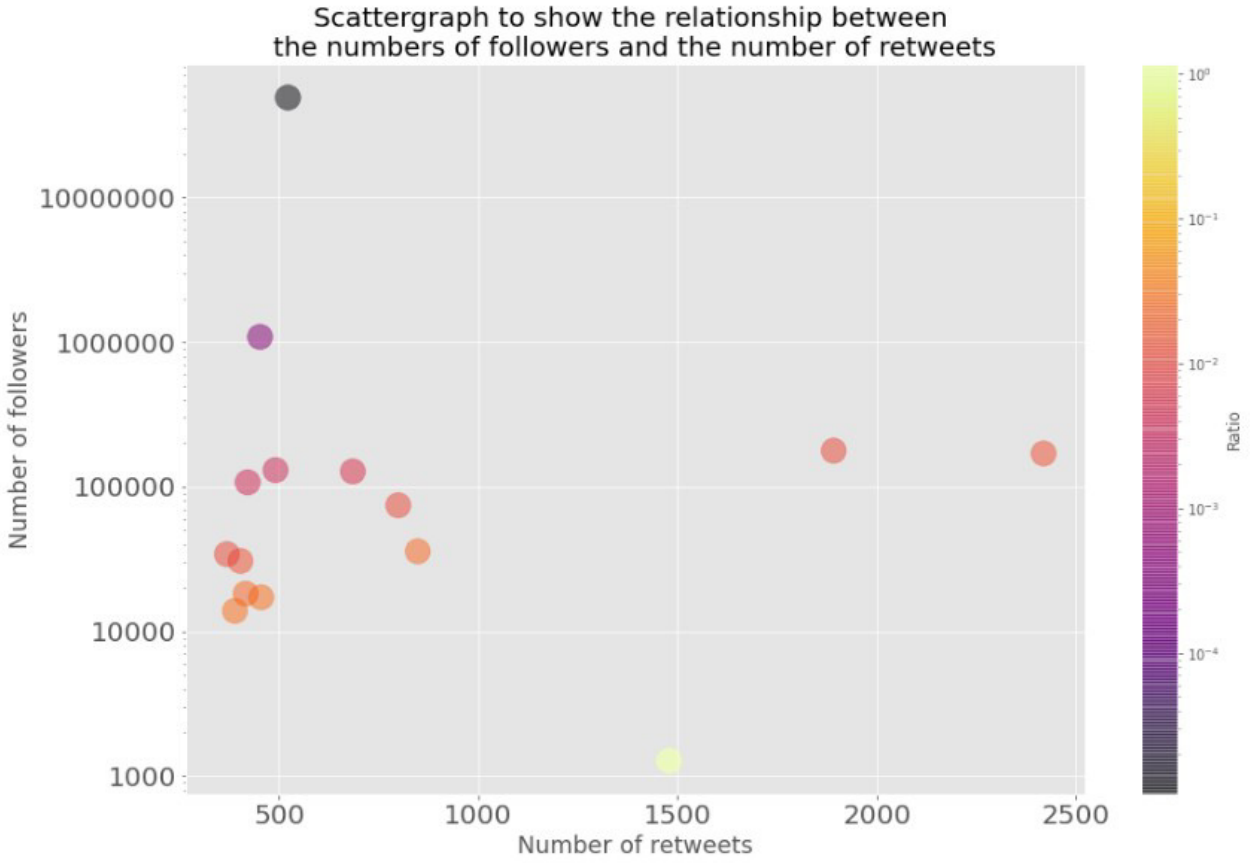


Figure 4.8: Scatter-graph of followers and retweets

Figure 4.9 shows the five most retweeted accounts in the anti-vaccination dataset. The top three accounts, @Jordan_Sather_, @cjtruth and @shoptaraeveland, have all tweeted anti-vaccination messages repeatedly, and spread conspiracy theories relating to “big pharma” and the “deep state”. The majority of all their retweets have come from right-wing affiliated accounts. With the retweets from left-wing accounts being so minuscule, it could be argued that a large proportion of the “Ambiguous” accounts are likely right-wing too. The account, @ChildrensHD, is the advocacy organisation, Children’s Health Defence, which are against vaccinations. It was founded by Robert F Kennedy, Jr., a prominent anti-vaxxer (English 2020). More ambiguous accounts retweeted @ChildrensHD, however, the right-wing accounts still retweeted the account at a much higher rate than left-wing accounts.

In fact the only account in the top five in which left-wing accounts retweeted an account more than right-wing accounts, was @PeterHotez. This is surprising, as the other four accounts are retweeted by right-wingers at a considerably higher rate. The account belongs to Peter Hotez, an American

vaccine scientist and paediatrician who is the Director of the Texas Children's Center for Vaccine Development. He has also personally led efforts to defend vaccines against the growing vaccine movement (Hotez 2020). Hotez had a total of 15 tweets in the data, all of them possessing a pro-vaccination message.

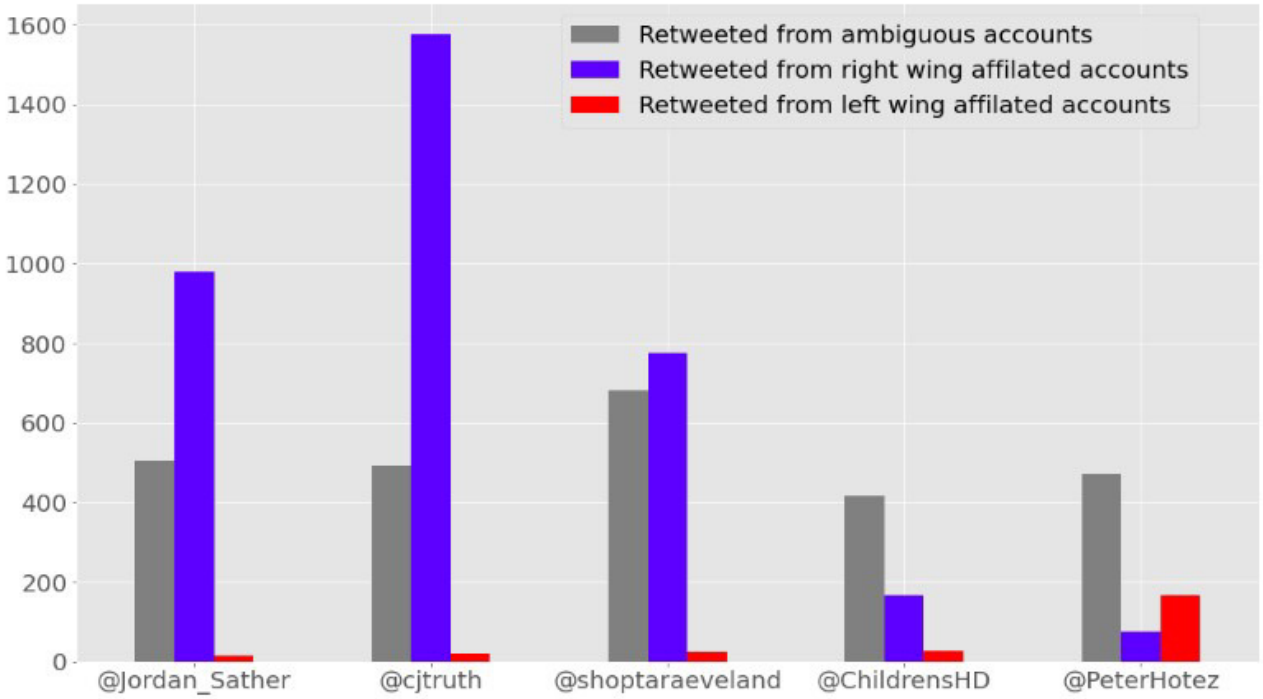


Figure 4.9: Top 5 most retweeted accounts

Tweets which express anti-vaccination sentiment or promote a conspiracy theory concerning vaccines seem to gain far more retweets than tweets promoting the necessity of vaccines. In addition the results again clearly suggest that the anti-vaccination movement is being driven by those on the right. They are much more likely to retweet anti-vaccination accounts and conspiracy theories related to vaccines thereby promoting and spreading them.

N-grams

N-grams are a concept that are found in Natural Language Processing. Essentially, they are a set of co-occurring words (Ganesan 2020). If N=2, which are known as bigrams, the n-grams of the following sentence would be:

"the quick brown fox jumped over the lazy dog"

- the quick

- quick brown
- brown fox
- fox jumped
- jumped over
- over the
- the lazy
- lazy dog

Meanwhile, if N=3, known as trigrams, the n-grams of the same sentence would be:

- the quick brown
- quick brown fox
- brown fox jumped
- fox jumped over
- jumped over the
- over the lazy
- the lazy dog

Being able to visualise the frequency of these n-grams should give a better understanding of the vaccination data, as well as what co-occurring words the left and right-wing accounts use.

In order to do this, the text will need to be cleaned in the same way the Twitter descriptions were in Chapter 1, i.e. remove all stop-words, put all text to lower case and stem words where it is necessary. The following vaccine-related words are added to the list of stop-words, as these words are the vaccine search terms used to identity vaccine-related tweets, meaning they would dominate the results.

```
Additional_stop_words = [ 'anti-vax', "anti-vaccincation", "anti vax", "vaccines","vaccine",
                          "vaccination", "vaccinations",'antivaxxer', 'antivaxxers', 'big pharma',
                          'pharmaceutical','antivaccine','rt', 'part', 'inject'] ]
```

Figure 5.0 shows the bigrams and trigrams of the overall vaccination dataset. The bigrams “deep, state” and “big, pharma” are clearly a reference to the conspiracy theories deep state and big pharma respectively. It is hardly surprising that these are some of the most frequent bigrams as they were both some of the most frequent hashtags as well. The bigram, “bill gate”, occurs well over 2000 times, and is reference to the Microsoft founder turned philanthropist Bill Gates. Along with his wife, Melinda, Gates set up the Gates Foundation which has been responsible for implementing vaccines in the developing world, which has helped eradicate polio (Belluz 2015). However, despite this, conspiracy theories surround him. Gates has been accused of manufacturing the coronavirus in order to depopulate the world and insert microchips into people (BBC 2020). A survey conducted by YouGov and Yahoo news found 25% of adults in America and 44% of Republicans believe this (Chang 2020).

The other bigram which represents a name is Dr. Fauci, the director of the National Institute of Allergies, and one of the lead members of the Trump administration’s White House Coronavirus Task Force. There have been reports that there is growing tension between Fauci and the Trump administration regarding the latter’s decision to rapidly reopen the economy, leading to many on the right accusing Fauci of being a “deep state doctor” (Cohen 2020).

The trigrams in figure 5.0 all possess very similar frequencies. This is because they are all retweets from the tweet by the user @shoptaraeveland which was discussed earlier. As a result of this results have been greatly skewed.

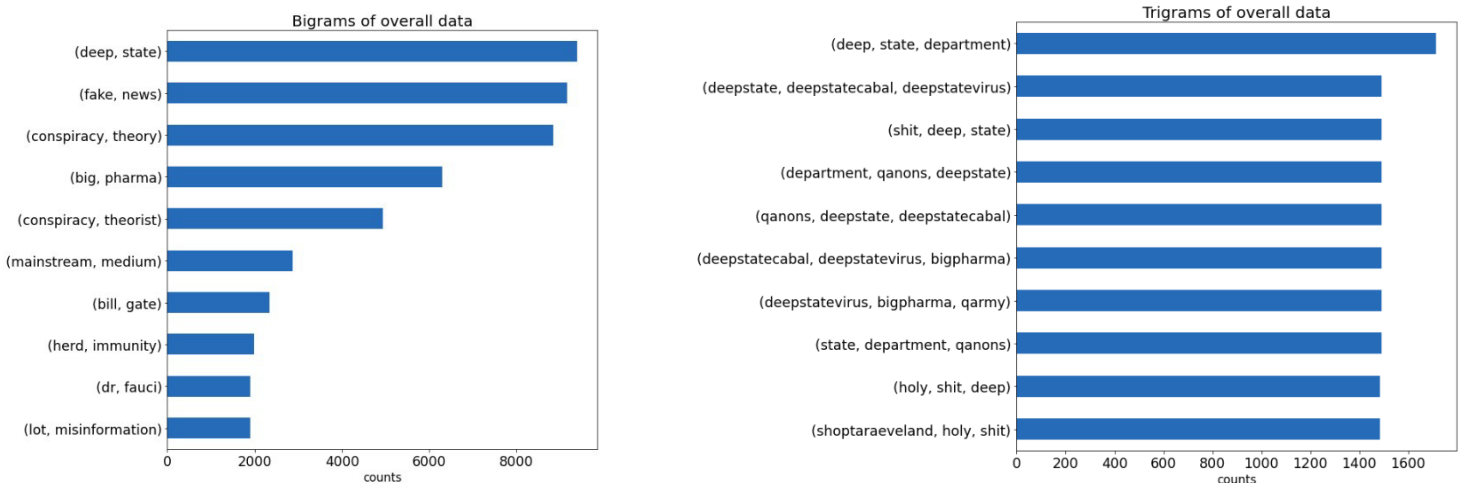


Figure 5.0: Bigrams and Trigrams

Figure 5.1 displays the most frequent bigrams for both the left and right-wing sets. The right-wing bigrams are fairly similar to the overall set with both containing the bigrams “deep state”, “big pharma” and “fake news”. The left-wing set has many matching bigrams with the overall data, “conspiracy theory”, “conspiracy theorist”, “fake news”, “deep state”, “big pharma” etc.

The bigrams, “fake news”, “deep state, “big pharma” are in both the top 10 of the left and right-wing most frequent bigrams. However they are much more frequent in the right-wing set, for instance, the bigram “deep state” appears over 4000 times, while it appears only around 400 times in the left-wing set.

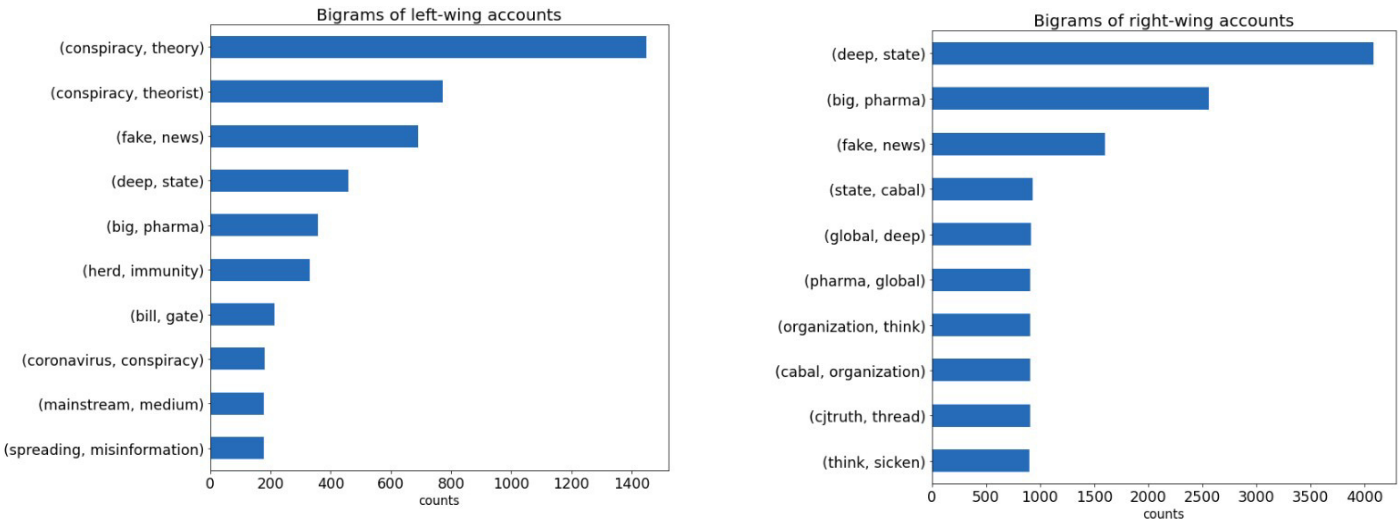


Figure 5.1: Bigrams of left/right-wing sets

Figure 5.1 shows the trigrams for both the left and right sets. Unfortunately, they have both been skewed by retweets. The left set has been been skewed by retweets from the account @rmayemsinger (the account is even listed in one of the trigrams). Meanwhile, the right has been skewed by retweets of the account @cjtruth. The reason for this is the limited amount of data for both sets. The left-wing set contains just over 10,000 tweets, while the right contains around 15,000. Therefore, a tweet with thousands of retweets in the data will greatly affect the results. Nonetheless, these results should not be ignored, as it further highlights what accounts the left and right are retweeting. The tweets of the two accounts are shown below:

- *@cjtruth- "Big Pharma is just a part of the global Deep State Cabal as any organization. I think it will sicken people when they find out what they have kept hidden and how truly corrupt they are."*
- @rmayemsinger - "*But it's not true, so the doctors forced him to walk it back. We've gone from fake news to fake vaccines. America, this is not okay*." (In response to a White House official saying doctors Trump met had developed a Covid-19 vaccine within three days in March).

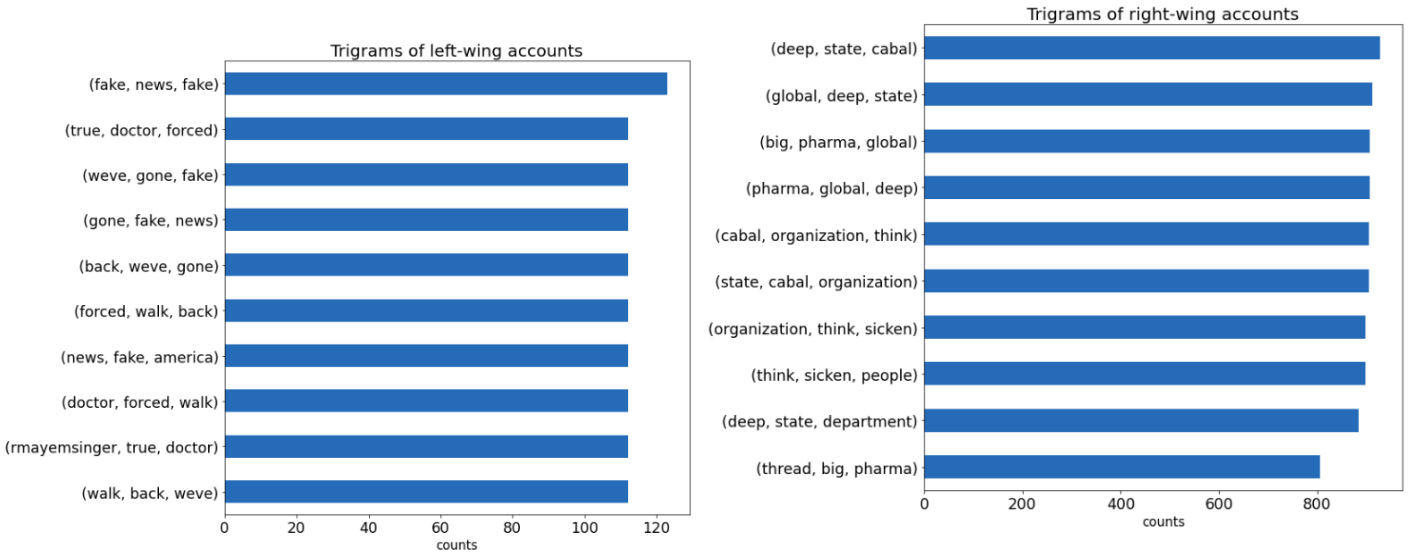


Figure 5.1: Trigrams of left/right-wing sets

Topic Modelling

Topic modelling is a form of statistical modelling for discovering the abstract "topics" that occur in a collection of documents (Li 2018). It is an unsupervised form of machine learning as it does not need a list of predefined list of tags that has been classified by humans (Monkey Learn 2020).

The "topics" that are produced by the topic modelling techniques are clusters of similar words. These clusters are captured by a mathematical framework, and based on the statistics of each word, topics are discovered along with the topic "weights". In this case the "documents" will be tweets.

Before the topic modelling can be applied, the text will have to be cleaned, just as the twitter descriptions before were in chapter 1. Links, users, punctuation, numbers, double spacing and

stopwords are all removed from the text, and the words are stemmed. The “cleaned tweets” are now a column called “stripped tweet”.

Figure 5.2 shows the text being turned into vector form. The remaining words will be filtered again to exclude words that appear in more than 90% of the text, and words that appear in less than 100 tweets will also be discarded. Words that appear in more 90% of the tweets are very unlikely to be relevant topics, while words that appear in less than 100 tweets will lead to an inefficiency in the program. In the code this is represented in the vectorizer object, max_df=0.9 (90%) and min_df=100.

The fit_transform method() transforms the the words in the column “stripped_tweet” into vector form. While the method, get_feature_names, gives what word in each column in the matrix represents.

```
from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer(max_df=0.9, min_df=100, token_pattern='\w+|\$[\d\.]+|\S+')

tf = vectorizer.fit_transform(overall_df['stripped_tweet']).toarray()

tf_feature_names = vectorizer.get_feature_names()
```

Figure 5.2: Vectorizer object

The model non-negative matrix factorisation (NMF) is being used for this study as opposed to the latent Dirichlet allocation. This is because the NMF model works better with shorter text documents, such as tweets (Github 2020).

The NMF model is a matrix factorisation method where the matrices are constrained to be non-negative. It approximates a matrix X with a low-rank such that X ≈ WH. It uses an iterative procedure to modify the values of W and H so that the product approaches X. This procedure will terminate after the specified number of iterations is reached. During the application of the model, the NMF model maps the data into a new set of “topics” discovered by the model (Oracle 2020) .

```
no_of_topics = 5

model = NMF(n_components=no_of_topics, random_state=0, alpha=.1, l1_ratio=.5, max_iter = 1000)
```

```
model.fit(tf)
```

Figure 5.3: How many topics

Figure 5.4 represents the topics for the overall vaccination dataset. Topic 0 contains the words “vaccin”, “misinform”,“bill”, “gate”, “covid”, “spread”. This is likely in reference to the misinformation spreading in the media regarding Bill Gates and vaccines.

Topic 1 is clearly in reference to the conspiracy theories deep state and big pharma. Meanwhile, in Topic 2 the highest weighted words are “fake” and “news”. The words “fraud”, “media”, “covid” and “exist” are present in Topic 2 as well. This suggests there is some doubt in some of the tweets as to whether Covid-19 exists. As mentioned before, there are conspiracies that the coronavirus pandemic has been state manufactured in order to implant “tracking devices” into the public.

	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights
0	vaccin	32.7	state	14.1	fake	16.5	conspiraci	24.4	...	19.4
1	misinform	3.5	deep	13.5	news	16.1	theori	11.7	propaganda	19.4
2	covid	2.6	big	9.9	big	3.7	theorist	5.2	immun	8.2
3	gate	1.9	pharma	9.6	fraud	3.1	...	3.4	inject	3.0
4	work	1.8	...	7.0	media	2.6	believ	2.0	misinform	2.8
5	media	1.6	they	3.5	covid	2.6	vaxxer	1.8	campaign	2.5
6	flu	1.6	part	3.1	immun	1.8	mask	1.6	lie	2.1
7	bill	1.4	cabal	2.9	de	1.8	world	1.5	doctor	1.9
8	spread	1.3	global	2.9	flu	1.8	immun	1.4	video	1.9
9	interfer	1.3	thread	2.8	exist	1.7	coronaviru	1.4	say	1.7

Figure 5.4: Topics for overall set

Figure 5.5 displays the topics for the right-wing set. The conspiracy “deep state” are the highest weighted words for four of the five topics, with the other highest weighted words being “big pharma” . Topic 2 and Topic 3 appear to be retweets from the accounts @shoptaraeveland and @Jordan_Sather_. It is clear the right-wing set is very conspiracy heavy, with all five topics containing at least one conspiracy theory.

	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights
0	state	5.4	vaccin	15.2	state	5.3	big	8.4	state	5.6
1	deep	5.4	conspiraci	4.6	deep	5.3	pharma	6.7	deep	5.6
2	pharma	5.0	propaganda	4.4	depart	4.2	…	5.7	…	4.8
3	big	4.8	…	3.3	shit	4.1	news	4.3	vaccin	4.3
4	peopl	4.5	amp	2.6	#bigpharma	4.1	amp	4.2	coronaviru	4.2
5	think	4.5	peopl	2.4	#qarmy	4.0	fake	4.1	say	4.2
6	…	4.4	viru	2.1	#deepstate	4.0	vs	3.7	wonder	3.9
7	part	4.4	theori	1.8	🙄	4.0	sather	3.1	mark	3.9
8	cabal	4.3	anti	1.6	#qanons	4.0	peopl	2.5	school	3.9
9	thread	4.3	covid	1.5	holi	4.0	attack	1.9	mouthpiec	3.9

Figure 5.5: Topics for right-wing set

Figure 5.6 shows the topics in the left-wing set. The words “misinform” and “disinform” are present in three out of the five topics which could suggest more of an awareness of misinformation/disinformation surrounding vaccinations. Also the words “deep state” are only present in Topic 0, with a significantly lower weighting than the right-wing set.

	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights
0	vaccin	15.6	immun	10.6	conspiraci	13.5	…	13.1	fake	8.7
1	misinform	1.8	propaganda	9.2	theori	6.4	misinform	5.0	news	6.7
2	covid	1.6	herd	1.3	theorist	2.6	danger	1.2	doctor	1.6
3	state	0.8	one	1.2	vaxxer	1.4	media	1.1	back	1.5
4	trump	0.8	system	1.0	coronaviru	1.3	#antivax	1.0	forc	1.4
5	gate	0.7	disinform	0.7	believ	1.2	spread	1.0	america	1.4
6	russia	0.7	public	0.7	mask	1.1	pharma	0.9	true	1.4
7	deep	0.7	lie	0.7	covid	0.9	inject	0.8	trump	1.0
8	make	0.6	hear	0.6	spread	0.8	ask	0.7	spread	0.7
9	test	0.6	know	0.5	big	0.8	vaxxer	0.7	coronaviru	0.6

Figure 5.5: Topics for left-wing set

Conspiracy theories surrounding vaccines are more far prevalent in the right-wing set than they are in left-wing set. These results support the earlier findings of the most frequent hashtags and retweeted accounts. The “deep state” and “big pharma” conspiracies hold very little weight in the left-wing topics, further suggesting they tend to be exclusive to right-wing accounts.

Verdict

Misinformation regarding vaccines on Twitter has increased exponentially since the start of the year. These significant increases are most likely due to the potential coronavirus vaccine. Much of this misinformation regarding vaccines has come under the form of conspiracy theories which range from the coronavirus being created by the “deep state” and “big pharma” to Bill Gates wanting to implant tracking devices into the population. These conspiracy theories do have a common theme, which is that they all concern the rich and powerful, be it governments, corporations or just extremely wealthy individuals, doing terrible things to the rest of the human population. Prof Joseph Uscinski, a political scientist at the University of Miami and author of books on conspiracy theories, explains, “*Conspiracy theories are about accusing powerful people of doing terrible things. The theories are basically the same”*. However, given the number of accounts retweeting these conspiracies it is clear people are believing them, or at the very least it is causing some to to have doubts around vaccinations.

In terms of the effect of political beliefs on vaccination misinformation, these conspiracy theories are being driven by right-wing accounts, who are more likely than the left-wing accounts to then retweet this misinformation. This was also evidenced by the fact that the overall vaccination set was being heavily influenced by the right-wing, in particular with the the most frequent hashtags and bigrams.

Chapter 4- Conclusions and Reflection

At the start of the project the overall goal was to establish what misinformation around vaccines is causing people to distrust them, and whether political leanings are having an effect on who promotes vaccine misinformation online. This goal has largely been met as it was fairly conclusive that conspiracy theories were the most frequent aspect of the vaccination dataset, and that it was mainly right-wing accounts promoting them.

The most significant achievement of this project was the left/right algorithm and its implementation. By using word correlations it was possible to validate and expand upon the list of left and right terms. In addition, the scoring system used to score strong/weak terms made minimal impact which illustrates the robustness of the algorithm.

The algorithm labelled accounts left or right-wing at a very high degree of accuracy, which was proven by who the left and right-wing accounts were retweeting most often. This allowed the analysis of both the left and right-wing sets.

Another achievement of the project was the use of topic modelling which gave an idea as to what subjects around vaccinations were being discussed. It also provided additional evidence that the right-wing are producing far more of the unproven conspiracies that surround vaccinations.

However, there were some deficiencies in the study, particularly in the analysis of the trigrams, where it was clear that there were not enough tweets in the datasets. All three of the trigrams analysed were skewed by a single tweet which had been retweeted numerous times. If there were more time, perhaps more data could have been collected. Additionally, the data had a very heavy bias towards America, which was shown by the most frequent words used in the Twitter descriptions. It would have been interesting to see the attitude of other countries towards vaccination, as well as their left and right split. However, this would have required many more different left/right terms in different languages, which would have been very difficult. It would also have been interesting to analyse online data other than Twitter to see how conclusions would vary.

Reflection

I learned a considerable amount throughout this project, with particular emphasis on machine learning techniques involving vectors which allowed the use of word correlations. I also gained an understanding of natural language processing, and successfully applied the NMF model to my data.

In terms of project management, I tried to split my time between conducting research and writing the actual report. I felt I struck a good balance between the two as I was not pressed for time when the deadline for the project was nearing. In addition my aims and objectives that I set at the start of the project kept me on track so my research was focussed and I largely avoided wasting unnecessary time.

One of the most important lessons I have taken from this project is to properly structure and organise code. For example, at the beginning of the project I made the error of using one notebook to conduct large amounts of my research. For instance, I had the left/right algorithm, hashtag analysis and total vaccine data all together in one notebook. This led to a great deal of confusion and inefficiency. To prevent this I split up my research making it easier to refer back to. I created a notebook which concerned only the left/right algorithm, another notebook for the hashtag research, and another solely devoted to topic modelling etc. This made my work much more organised and now I recognise the important of code organisation for such projects.

References

Alley, K. (2016). Kirstie Alley endorses Donald Trump. Arlington County. [Online]. Available at: https://www.politico.com/story/2016/04/kirstie-alley-endorses-donald-trump-221753 [Accessed 13 September 2020].

Andrews, E. (2015). [Online]. New York. Available at: https://www.history.com/news/how-did-the-political-labels-left-wing-and-right-wing-originate [Accessed 02 September 2020].

The Babylon Bee. (2020). About us. Jupiter. [Online]. Available at: **https://babylonbee.com/about** [Accessed 01 September 2020].

Baker, P. (2018). New York. 'Use That Word!': Trump Embraces the 'Nationalist' Label. [Online] Available at: **https://www.who.int/vaccine_safety/initiative/communication/network/vacctoday/en/** [Accessed 23 August 2020].

Ball, P. (2020). Anti-vaccine movement could undermine efforts to end coronavirus pandemic, researchers warn. [Online]. London. Available at: https://www.nature.com/articles/d41586-020-01423-4 [Accessed 15 July 2020].

BBC. (2016). Pepe the Frog meme branded a 'hate symbol'. London. [Online]. Available at: https://www.bbc.co.uk/news/world-us-canada-37493165 [Accessed 21 August 2020].

BBC. (2019). OK hand sign added to list of hate symbols. London. [Online]. Available at https://www.bbc.co.uk/news/newsbeat-49837898 [Accessed 26 August 2020].

BBC. (2020). Trump angers Beijing with "Chinese Virus" tweet. London. [Online]. Available at: https://www.bbc.co.uk/news/world-asia-india-51928011 [Accessed 17 September 2020].

Beinart, P. New York. Mike Pompeo's Allies on the Anti-Muslim Right. [Online]. Available at **https://www.theatlantic.com/international/archive/2018/03/pompeo-muslims/555680/** [Accessed 27 August 2020].

Belluz, J. (2015). The media loves the Gates Foundation. These experts are more skeptical. Washington D.C. [Online]. Available at: **https://www.vox.com/2015/6/10/8760199/gates-foundation-criticism** [Accessed on 21 September 2020].

Belluz, J. (2019). America is in danger of losing its "measles-free" status. [Online]. Washington, D.C. Available at: https://www.vox.com/2019/9/11/20850836/meales-outbreak-2019-cdc [Accessed 14 July 2020].

Berg, D. (2019). Left-wing activist groups celebrates 'international pronouns day'. [Online]. Available at: https://sovereignnations.com/2019/10/17/left-wing-activist-international-pronouns-day/ [Accessed 03 September 2020].

Blaskiewicz, R. (2013). The Big Pharma conspiracy theory.

BMJ. (2011). Wakefield's article linking MMR vaccine and autism was fraudulent. 342:c7452.

Bongino, *D. (2016)*. Dan Bongino Loses Florida Congressional Race. [Online]. Available at: https://scotteblog.com/2016/08/31/dan-bongino-loses-florida-congressional-race/ *[Accessed 16 September 2020]*.

Boseley, S. Coronavirus: fifth of people likely to refuse Covid vaccine, UK survey finds. London. Available at: *https://www.theguardian.com/world/2020/sep/24/a-fifth-of-people-likely-to-refuse-covid-vaccine-uk-survey-finds* [Accessed 12 September 2020].

thebradfordfile. (2020). *I don't know who needs to hear this but Trump doesn't negotiate with terrorists. Fill the seat.* [Twitter]. 19 September 2020. Available at: https://twitter.com/thebradfordfile/status/1307310795139813377 [Accessed: 21 September 2020].

FOOL NELSON. (2020). *RT #BREAKING Senate #Ukraine Report On #Biden, #Burisma Expected 'in days' Says Top #GOP Senator.* [Twitter]. 18 September 2020. Available at: https://twitter.com/SaraCarterDC/status/1306975141147090947 [Accessed: 22 September 2020].

Broniatowski, D,A. Jamison, A, M, Qi, S. AlKulaib, L. Chen, T. Benton, A. Quinn, S, C. Dredze, M. Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. 1378–1384.

Catturd. 2020. *I'm completely shocked and stunned every single day that the Democrats keep Joe Biden as their nominee. This is the biggest train wreck in U.S. political history.* [Twitter]. 21 Septmeber. Available at: https://twitter.com/catturd2/status/1307821310609895425 [Accessed: 21 September 2020].

CBS News. (2020). What is the QAnon conspiracy theory? New York. [Online]. Available at: https://www.cbsnews.com/news/what-is-the-qanon-conspiracy-theory/ [Accessed 29 August 2020].

Chang, B. (2020). More than 40% of Republicans in a new poll say they think Bill Gates wants to use COVID-19 vaccines to implant location-tracking microchips in recipients. New York. [Online]. Available at: https://www.businessinsider.com/republicans-bill-gates-covid-19-vaccine-tracking-microchip-study-2020-5?r=US&IR=T [Accessed on 20 September 2020].

Chang, J., Riegle, A., Effron, L. (2019). 'Star Trek' star George Takei on why his activism roots are deeply personal and being a Twitter legend. New York. [Online]. Available at: https://abcnews.go.com/Entertainment/star-trek-star-george-takei-activism-roots-deeply/story?id=64932887 [Accessed 29 August 2020].

Cillizza, C. (2020). Fox's firing of Diamond & Silk isn't the problem. Fox's hiring of them is. Atlanta. [Online]. Available at: https://edition.cnn.com/2020/04/28/politics/diamond-silk-fox-news-donald-trump/index.html [Accessed 14 September 2020].

Cohen, S. (2020). Is Fauci A "Deep State" Doctor? The Conspiracy Theory That Is Sickening America. Jersey City. [Online]. Available at: https://www.forbes.com/sites/sethcohen/2020/07/14/is-fauci-a-deep-state-doctor/#c6fc7d272550 [Accessed 20 September 2020].

Coppins, M. (2018). The Man who broke politics. New York. [Online]. Available at: https://www.theatlantic.com/magazine/archive/2018/11/newt-gingrich-says-youre-welcome/570832/ [Accessed 02 September 2020].

Dar, P. (2018). Comprehensive Beginner's Guide to Jupyter Notebooks for Data Science & Machine Learning. [Online]. Available at: https://www.analyticsvidhya.com/blog/2018/05/starters-guide-jupyter-notebook/ [Accessed 08 September 2020].

Dictionary. (2020). Cross Mark emoji. [Online] Available at: https://www.dictionary.com/e/emoji/cross-mark-emoji/ *[Accessed 25th August 2020]*.

Dictionary. (2020). Deep State. [Online] Available at: https://www.dictionary.com/e/politics/deep-state/ [Accessed 19 September 2020].

Dizikes, P. (2018). Study: On Twitter, false news travels faster than true stories. [Online]. Cambridge, Massachusetts. Available at: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308 [Accessed 16 July 2020].

Dworkin, S. (2020). Scott Dworkin. [Online]. Available at: https://scottdworkin.org/ [Accessed 15 September 2020].

Eichler, W. (2015). Singling out Israel: a perspective from the left. [Online]. Available at: https://www.opendemocracy.net/en/north-africa-west-asia/singling-out-israel-perspective-from-left/ [Accessed 05 September 2020].

English, C. (2020). Anti-Vaccine Group 'Children's Health Defense' Smells A Coronavirus Conspiracy. New York. [Online]. Available at https://www.acsh.org/news/2020/04/01/anti-vaccine-group-childrens-health-defense-smells-coronavirus-conspiracy-14681 [Accessed at 20 September 2020].

FBI. (2011). C. Frank Figliuzzi Appointed as Assistant Director of the FBI's Counterintelligence Division. Washington D.C. [Online]. Available at https://archives.fbi.gov/archives/news/pressrel/press-releases/c.-frank-figliuzzi-appointed-as-assistant-director-of-the-fbi2019s-counterintelligence-division [Accessed 15 September 2020].

Finnegan, G. (2019). Do our political views influence vaccination rates? [Online]. Available at: **https://www.vaccinestoday.eu/stories/do-our-political-views-influence-vaccination-rates/** [Accessed 16 September 2020].

Flood, B. (2020). Democrats' go-to mainstream media outlets having a hard time saying anything nice. New York [Online]. Available at: **https://www.foxnews.com/media/democrats-mainstream-media-troubles** [Accessed 29 August 2020].

Folley, A. (2020). [Online]. Available at: **https://www.msn.com/en-us/news/politics/james-woods-defends-trump-he-loves-america-more-than-any-president-in-my-lifetime/ar-BB14dOHE** [Accessed on 20 August 2020].

Ganesan, K. (2020). What are N-Grams? [Online]. Available at: **https://kavita-ganesan.com/what-are-n-grams/#.X3BZpnVKg5l** [Accessed at 20 September 2020].

Gillin, J. (2015). St. Petersburg. Jeb Bush Says Donald Trump "was a Democrat longer in the last decade than he was a Republican." [Online]. Available at: **https://www.politifact.com/factchecks/2015/aug/24/jeb-bush/bush-says-trump-was-democrat-longer-republican-las/** [Accessed 24 August 2020].

GitHub. 2020. Topic Modelling in Python. [Online]. Available at: **https://ourcodingclub.github.io/tutorials/topic-modelling-python/** [Accessed 26 July 2020].

GP Online. (2010). MMR doctor Andrew Wakefield struck off by GMC for misconduct. [Online]. Available at: **https://www.gponline.com/mmr-doctor-andrew-wakefield-struck-off-gmc-misconduct/article/1005151** [Accessed: 12 July 2020].

GrimKim. 2020. *Union-busting at progressive mission-focused nonprofit workplaces is still union-busting!!!* [Twitter]. 21 September 2020. Available at: **https://twitter.com/GrimKim/status/1308165147979067403** [Accessed 23 September 2020].

Hayden, M. (2020). Jack Posobiec's Rise Tied to White Supremacist Movement. Montgomery. [Online]. Available at: **https://www.splcenter.org/hatewatch/2020/07/08/jack-posobiecs-rise-tied-white-supremacist-movement** [Accessed 15 September 2020].

Huffington Post. (2020). Chris Evans On Gay Marriage: 'In 10 Years We'll Be Ashamed That This Was An Issue'. New York. [Online]. Available at: **https://www.huffpost.com/entry/chris-evans-on-gay-marriage_n_1442704** [Accessed at 16 September 2020].

Hussain, A., Ali, S., Ahmed, M., Hussain, S. (2018). The Anti-vaccination Movement: A Regression in Modern Medicine.

Igoe, K J. (2019). Establishing the Truth: Vaccines, Social Media, and the Spread of Misinformation. [Online]. Cambridge, Massachusetts. Available at: https://www.hsph.harvard.edu/ecpe/vaccines-social-media-spread-misinformation/ [Accessed at 14 July 2020].

Illing, S. (2020). The "deep state" is real. But it's not what Trump thinks it is. Washington D.C. [Online]. Available at: https://www.vox.com/policy-and-politics/2020/5/13/21219164/trump-deep-state-fbi-cia-david-rohde [Accessed 19 September 2020].

Jabeen, H. (2018). Stemming and Lemmatization in Python. [Online]. Available at: https://www.datacamp.com/community/tutorials/stemming-lemmatization-python [Accessed 15 September 2020]. Bandung, pp 160-167.

Jatnika, D. Bijaksana, M. Suryani, A. (2019). Word2Vec Model Analysis for Semantic Similarities in English Words. 4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12-13 September 2019.

Jordan_Sather. 2020. *The "Really American PAC" put this slick propaganda attack vs. QAnon. It's been retweeted enough (probably Jack's twitter bots) that.* [Twitter] 21 September 2020. Available at: https://twitter.com/Jordan_Sather_/status/1308160161106726913 [Accessed 22 September 2020].

Li, S. (2018). Topic Modeling and Latent Dirichlet Allocation (LDA) in Python. [Online]. Available at: https://towardsdatascience.com/t.opic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24 [Accessed 22 September 2020].

Lopez, G. (2016). Does the scientific community support vaccination? [Online]. Washington, D.C. Available at: https://www.vox.com/2018/8/21/17588092/vaccines-science-community-evidence [Accessed 14 July 2020].

Matplotlib. (2020). Matplotlib: Visualization with Python. [Online]. Available at: https://matplotlib.org/ [Accessed 10 September 2020].

Mazzetti, M. Savage ,C. Goldman, A. (2020). How Michael Flynn's Defense Team Found Powerful Allies. New York. [Online]. Available at: https://www.nytimes.com/2020/06/28/us/politics/michael-flynn-sidney-powell.html [Accessed 14 September 2020].

McCandless, D & Posavec, S. (2010). Left vs. Right (US). [Online]. Available at: https://www.informationisbeautiful.net/visualizations/left-vs-right-us/ [Accessed 02 September 2020].

McCoy, C. (2017). Anti-vaccination beliefs don't follow the usual political polarization. [Online]. Available at: **https://theconversation.com/anti-vaccination-beliefs-dont-follow-the-usual-political-polarization-81001** [Accessed 16 September 2020].

Mercia, D. Tatum, S. Clinton expresses regret for saying 'half' of Trump supporters are 'deplorables'. Atlanta. [Online]. Available at: https://edition.cnn.com/2016/09/09/politics/hillary-clinton-donald-trump-basket-of-deplorables/index.html [Accessed 31 August 2020].

Morales, V. (2012). [Online]. Washington D.C. Available at: https://www.voanews.com/usa/us-politics/experts-explain-absence-socialism-america [Accessed 24 August 2020].

Moran, L. (2020). Supercut Exposes The Ugly Truth Of Donald Trump's Rhetoric On Protests. New York. [Online]. Available at: https://www.huffingtonpost.co.uk/entry/donald-trump-ugly-presidency-meidas-touch_n_5ee332c9c5b6b13c3bdb5f23?ri18n=true [Accessed at 16 September 2020].

MSNBC. (2020). Frank Figliuzzi: Greatest threat to the country is an 'insider threat sitting inside the Oval Office'. New York. [Online]. Available at: https://www.msnbc.com/deadline-white-house/watch/frank-figliuzzi-greatest-threat-to-the-country-is-an-insider-threat-sitting-inside-the-oval-office-90270277986 [Accessed 15 September 2020].

NBC news. (2016). Trump and Other Conservatives Embrace 'Blue Lives Matter' Movement. New York. [Online]. Available at: https://www.nbcnews.com/politics/2016-election/trump-other-conservatives-embrace-blue-lives-matter-movement-n615156 [Accessed at 15 September 2020].

Nelson, R. (2015). The 21-Year-Old Becoming a Major Player in Conservative Politics. New York. [Online]. Available at: https://www.theatlantic.com/politics/archive/2015/03/the-21-year-old-becoming-a-major-player-in-conservative-politics/451110/ [Accessed 15 September 2020].

The New York Times. (2016). Election 2016: Exit Polls. New York [Online]. Available at: https://www.nytimes.com/interactive/2016/11/08/us/politics/election-exit-polls.html [Accessed 25 August 2020].

Nltk. (2020). Natural Language Toolkit. [Online]. Available at: https://www.nltk.org/ [Accessed 10 September 2020].

NHS. (2019). 'No link between MMR and autism,' major study finds. [Online]. London. Available at: https://www.nhs.uk/news/medication/no-link-between-mmr-and-autism-major-study-finds/ [Accessed 13 July 2020].

Oldham, S. (2020). Chris Evans Blasts Trump's Response to Pandemic: 'America Wants Leadership'. Santa Monica. Available at: https://variety.com/2020/film/news/chris-evans-trump-coronavirus-marvel-president-1203535209/ [Accessed at 16 September 2020].

On the Issues. (2017). 2016 Republican nominee for President; 2000 Reform Primary Challenger for President. [Online] Available at: https://web.archive.org/web/20171004000455/http://www.ontheissues.org/Donald_Trump.htm [Accessed 25 September 2020].

Oracle. (2020). Non-Negative Matrix Factorization. Redwood City. [Online]. Available at: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/algo_nmf.htm#DMCON058 [Accessed on 22 September 2020].

pant_leg 2020. *woman *has a body* everyone in the world for some reason.* [Twitter] 14 September 2020. Available at: https://twitter.com/pant_leg/status/1305596410335768576 Accessed 23 September 2020].

Pimlott N. (2019). Vaccine hesitancy and the art of family medicine. *Can Fam Physician*.

PyMOTW. (2020). re – Regular Expressions. [Online]. Available at: https://pymotw.com/2/re/ [Accessed 10 September 2020].

Rans, B. (2008). Remembering Harriet Tubman in Today's World. [Online]. Available at: https://progressive.org/op-eds/remembering-harriet-tubman-today-s-world/ [Accessed 14 September 2020].

Rao, T. S. Sathyanarayana., Andrade, C. (2011). The MMR vaccine and autism: Sensation, refutation, retraction, and fraud.

Reuters, Sarah. (2017). FBI says police deaths spiked 61% in 2016. New York. [Online]. Available at: https://www.businessinsider.com/r-us-police-deaths-on-duty-spiked-in-2016-fbi-2017-10?r=US&IR=T [Accessed 15 September 2020].

Revesz, R. (2016). Donald Trump's son spearheads his presidential fundraising campaign. London. [Online]. Available at: https://www.independent.co.uk/news/world/americas/us-politics/donald-trump-eric-trump-son-spearheads-fundraising-campaign-seeks-10-million-a7106366.html [Accessed 14 September 2020].

Science Daily. (2020). Vaccine misinformation and social media. [Online]. Rockville. Available at: https://www.sciencedaily.com/releases/2020/02/200217163004.htm [Accessed 16 September 2020].

Schaffer, A. (2019). Fear, Misinformation, and Measles Spread in Brooklyn. [Online]. San Francisco. Available at: https://www.wired.com/story/fear-misinformation-measles-spread-in-brooklyn/ [Accessed 14 July 2020].

Seaborn. (2020). Seaborn. Available at: https://seaborn.pydata.org/#:~:text=Seaborn%20is%20a%20Python%20data,attractive%20and%20informative%20statistical%20graphics. [Accessed 10 September 2020].

Slavitt, A (2020). Andy Slavitt. [Online]. Available at: https://medium.com/@ASlavitt [Accessed 10 September 2020].

Srivastava, S. (2019). Top 10 Data Science Programming Languages For 2020. [Online]. Available at: https://www.analyticsinsight.net/top-10-data-science-programming-languages-for-2020/ [Accessed 02 September 2020].

Stat. (2019). It’s old news that vaccines don’t cause autism. But a major new study aims to refute skeptics again. [Online]. Boston. Available at: https://www.statnews.com/2019/03/04/vaccines-no-association-autism-major-study/ [Accessed 13 July 2020].

Stelter, B. (2017). Birth of a conspiracy theory: How Trump's wiretap claim got started. Atlanta. [Online]. Available at: https://money.cnn.com/2017/03/06/media/mark-levin-joel-pollak-breitbart-trump-obama/index.html. [Accessed 15 September 2020].

Swearingen, J (2020). Elon Musk tweets 'take the red pill' in another strange turn for the billionaire. New York. [Online]. Available at: https://www.businessinsider.com/elon-musk-tweets-take-the-red-pill-what-it-means-2020-5?r=US&IR=T [Accessed 26 August 2020].

Tribe, L. (2019). The House must flex its constitutional muscles to get to Trump. London. [Online] Available at https://www.theguardian.com/commentisfree/2019/sep/30/the-house-constitutional-trump-impeachment-process-president [Accessed 30 August 2020].

Tutorials Point. (2020). [Online]. Available at: https://www.tutorialspoint.com/python_text_processing/python_tokenization.htm#:~:text=In%20Python%20tokenization%20basically%20refers,in%20programs%20as%20shown%20below. [Accessed 12 September 2020].

Tweepy. (2020). Tweepy. [Online]. Available at: https://www.tweepy.org/ [Accessed 10 September 2020].

Twitter. (no date given). How to use hashtags. San Francisco. [Online]. Available at: https://help.twitter.com/en/using-twitter/how-to-use-hashtags#:~:text=A%20hashtag%E2%80%94written%20with%20a,topics%20they%20are%20interested%20in. [Accessed at 17 September 2020].

Twitter. (no date given). Retweet FAQs. San Francisco. [Online]. Available at: https://help.twitter.com/en/using-twitter/retweet-faqs [Accessed at 17 September 2020].

Wakefield, J. (2020). How Bill Gates became the voodoo doll of Covid conspiracies. London. [Online]. Available at: https://www.bbc.co.uk/news/technology-52833706 [Accessed on 20 September 2020].

Watson, I. (2020). Unite election: Battle to succeed Len McCluskey heats up. London. [Online]. Available at: https://www.bbc.co.uk/news/uk-politics-53832653 [Accessed 13 September 2020].

Wong, J. (2020). QAnon explained: the antisemitic conspiracy theory gaining traction around the world. London [Online]. Available at: https://www.theguardian.com/us-news/2020/aug/25/qanon-conspiracy-theory-explained-trump-what-is [Accessed 29 August 2020].

Woods, J. 2020. *Reading the "news" is a soul-crushing exercise in pure misery. The mainstream media have slowly, silently degraded journalism into sheer bald-faced propaganda with stealth and guile. They are like diseased weasels prowling in the night.* [Twitter] 13 July. Available at: https://twitter.com/realjameswoods/status/1282503512409010176?lang=en [Accessed: 28 August 2020].

World Health Organisation. (2020). Geneva. Vaccines Today. [Online]. Available at: https://www.who.int/vaccine_safety/initiative/communication/network/vacctoday/en/ [Accessed 16 September 2020].

Wulfsohn, J. (2020). Twitter apologizes after briefly suspending The Babylon Bee's account. New York. [Online]. Available at: https://www.foxnews.com/media/twitter-briefly-suspends-babylon-bee-then-apologizes [Accessed 01 September 2020].