November 21, 2021

# News Media Companies Tweet Analysis

## Sentiment and Text Analysis

Usama Tariq

# Table of Contents



# Table of Figures

# Introduction:

Data has grown at an exponential rate in the past decade. The total amount of data in the world was 4.4 zettabytes in 2013. That has risen steeply to 44 zettabytes by 2020. To put that in perspective, 44 zettabytes is equivalent to 44 trillion gigabytes and now when social media has become an integral part of our lives, a major part of this data is generated by them. People from all over the world use these platforms to express their emotions and share their views and opinions about everything and anything.

Millions of people using these platforms generate sentiment rich data in the form of tweets, posts, videos, status updates, reviews, comments etc. Through it we observe how trends are generated and people and communities all over the world are influenced. This provides an interesting opportunity to various Government organizations, marketing firms, businesses and research institutes to study people and generate actionable intelligence which can be used to produce legislations, advertising campaigns, new products and services, academic research etc.

In this report we aim to highlight how we can analyze the sentiment rich data from Twitter tweets about five News/Media companies. By studying the sentiment generated by each tweet we can observe patterns and trends which can tell us more about the company's social media presence and how it effects the company and industry eco-system.

# Methodology:

## Project Description:

The whole purpose of this project was to study the tweets of five News/Media companies and study the sentiment produced by each tweet. We also want to observe how their tweets influence the industry, people, and world in general.

We used R's "TwitteR" package to scrap tweets from Twitter, store them in a data frame. Then using a GitHub package called "sentiment", we parse through the tweets and assign each tweet with a sentiment of either, positive, negative, or neutral. In the end we clean the tweets to remove unnecessary stop words and generate plots and graphs which can help us visualize the frequently used words and our sentiment analysis.

## Data Description:

The Data for this project was scarped in form of tweets from Twitter using the twitter developer account and R's "TwitteR" package. Along with the tweets we were able to get more information like the date they were created, if they were retweeted, the device on which the tweet was generated and if the tweet was a reply to another twitter account or not.

The News/Media Outlets chosen were:

1. CNN Breaking News (@cnnbrk)
2. New York Times (@nytimes)
3. BBC Breaking News (@BBCBreaking)
4. The Economist (@TheEconomist)
5. Reuters (@Reuters)

We specifically specified in our code to download 1000 tweets of each account. However, with a hobbyist Twitter Developer account you can just download tweets of the past week. This led to us getting tweets less than 1000. We got a total of 3425 tweets out of which CNN Breaking News had

812, New York Times had 221, BBC Breaking News had 496, The Economist had 906 and Reuters had 990 tweets.

### Data Arrangement:

After downloading the tweets from each twitter account, we converted them to a data frame. Next, we merged the 5 data frames into a single data frame using "rbind" function. This gave us a single data frame in which all the required information to run the sentiment and text analysis was aggregated in the columns.

## Sentiment Analysis:

Using the now arranged data frame we ran the sentiment analysis, using the "sentiment" function of the sentiment package we downloaded from GitHub. The package has inbuilt lists of words which it has categorized as positive, negative, or neutral. The function parses through each tweet and compares the words of the tweets and the lists, and based on the number of positive, negative, or neutral words in the tweets it assigns the tweet a polarity score and status. 4 means positive, 2 means neutral and 0 means negative.

Once we have the polarity of each tweet, we assign it to our data frame so that we can use it to run our text analysis. Lastly, we export the data frame as an excel file using the "write_xlsx" function of "writexl" package.

## Text Analysis:

To perform text analysis on our data frame we first need to convert it into a corpus. This converts the data frame into a large, structured set of text. After we have our corpus we tokenize the text, this allows us to manipulate the text e.g., we can remove stop words, punctuation, symbols etc. from our text and make it cleaner for our text analysis.

In the next step we convert our now clean corpus into a document feature matrix, which has documents in rows and terms in columns. This matrix tells us about the frequency of terms that occur in our documents. Using this matrix, we can obtain tables and visualizations depicting words/terms most frequently used in our tweets. Based on the context of the tweets and events behind and surrounding these tweets we can assign meaning to the frequency of words and drive insights from them.

## Visualizations:

### Sentiment Analysis Visuals:

For the Sentiment analysis part of our project, we use the excel file generated to plot three graphs. The first being a bar graph showing the overall distribution of polarity among all the tweets, the second being a pie chart that describes the percentage division of the polarity among the tweets and the third being the distribution of polarity among the tweets subcategorized by the companies. Shown in Appendix A

### Insights – Sentiment Analysis Visuals:

We can see from the visuals that majority of the tweets are of neutral sentiment, 3144 in total amounting to 92% of the total. In comparison the tweets with positive and negative sentiment are considerably less, 120 and 161 respectively. They only amount to 3% and 5% of the total. This is not surprising as these are Media outlets and most of their tweets are news related as such it's their fundamental duty to report news in an unbiased manner. This leads to a lot of neutral sentiment tweets

### Text Analysis Visuals:

In the text analysis part of the project, we aim to depict the use of words and their frequency in our document feature matrix. This allows us to explain the context in which they were used and assign meaning to them. In this regard we used a table containing all the frequent words in our document feature matrix, a packed bubble graph showing the top ten most frequently used words and three-word clouds depicting the words, sized by their frequency in the tweets. Shown in Appendix B

### Insights – Text Analysis Visuals:

From the visuals we can observe that the most frequently used words are “people”, “president”, “covid-19”, “police”, “former”, “climate”, “vaccine”, “government” and “rittenhouse”. The reason behind that is that each of the word corresponds to events that were trending in the past week. Those events being:

1. President Biden saying that people need to respect the court’s decision in the Kyle Rittenhouse case.
2. President Biden relinquishing power to VP Harris because of his colonoscopy.
3. The Kyle Rittenhouse trail.
4. The Covid-19 pandemic and booster shot of vaccine.
5. Recent extreme changes in climate because of Global warming.
6. The military show of power of China and its allies and US and its allies in case of Chinese threat of invasion to Taiwan.

Using the above events as context we can explain the frequent use of the words.

## Conclusions:

In this day and age social media sites are powerful tools that can be used to bring change and influence. In this regard Twitter is considered as the best online source of news and trends. If used effectively it can bring about a positive impact on society. Which is how most companies, no matter which industry they belong to, strive to use it.

In case of our chosen companies which belong to the News and Media industry we can say that they have used their online presence for this purpose effectively. By studying the tweets, tables and graphs we can conclude that:

1. By studying the tweet polarity distribution, we can say that as Media outlets they have remained unbiased in their reporting of the News. As most of the tweets have been carefully curated with neutral words.
2. As it’s the week of Thanksgiving we see tweets of positive polarity.
3. Due to recent events like US-China tension, Kyle Rittenhouse case, Covid-19 pandemic we see tweets of negative nature.
4. The overall percentage of positive and negative tweets, 3% and 5% respectively can be explained by the fact that we are studying Media outlets and due to their word’s influence, they must report all sort of news in an especially careful manner.
5. Lastly, as these five media outlets have the most amount of twitter followers for News companies, we can correctly assume that they have influenced the public. This can be observed by the recent viral hashtags on Twitter which contained all our analysis’s frequently used words.

# Appendix A:

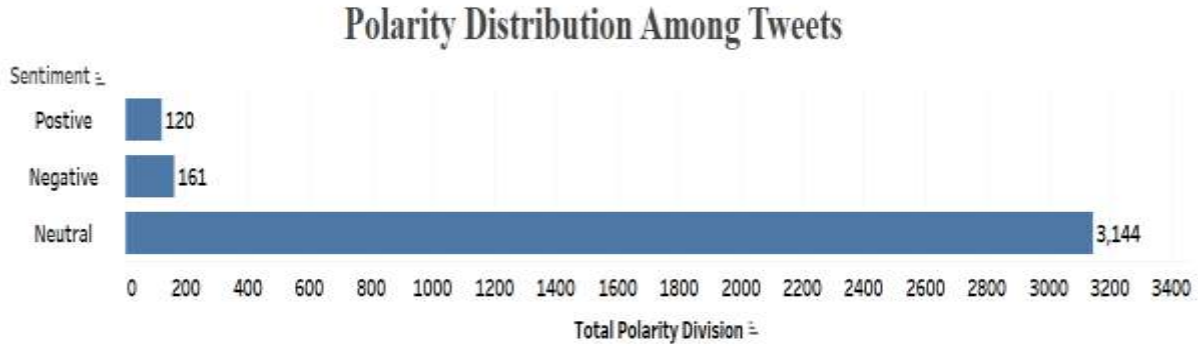


*Figure 1 The Total Polarity Distribution among tweets*

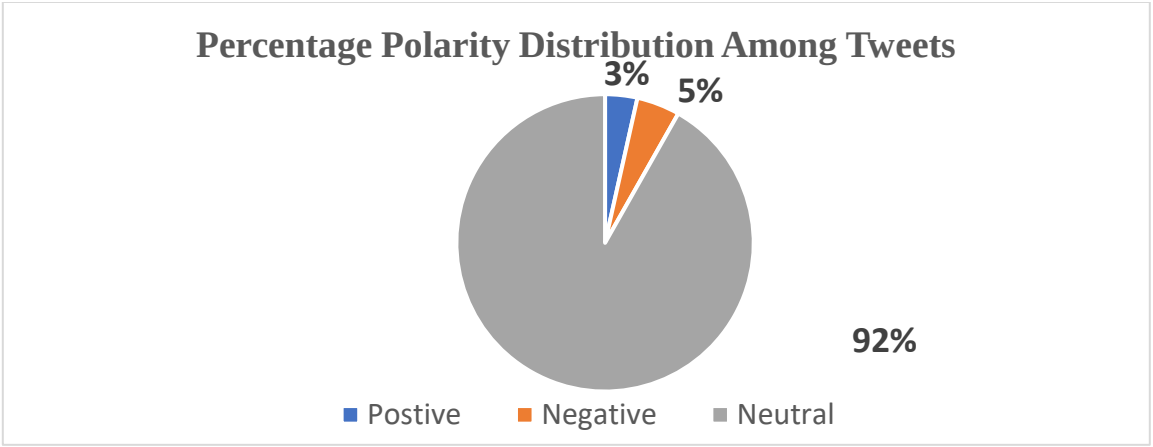


*Figure 2 The Percentage Total Polarity Distribution among tweets*

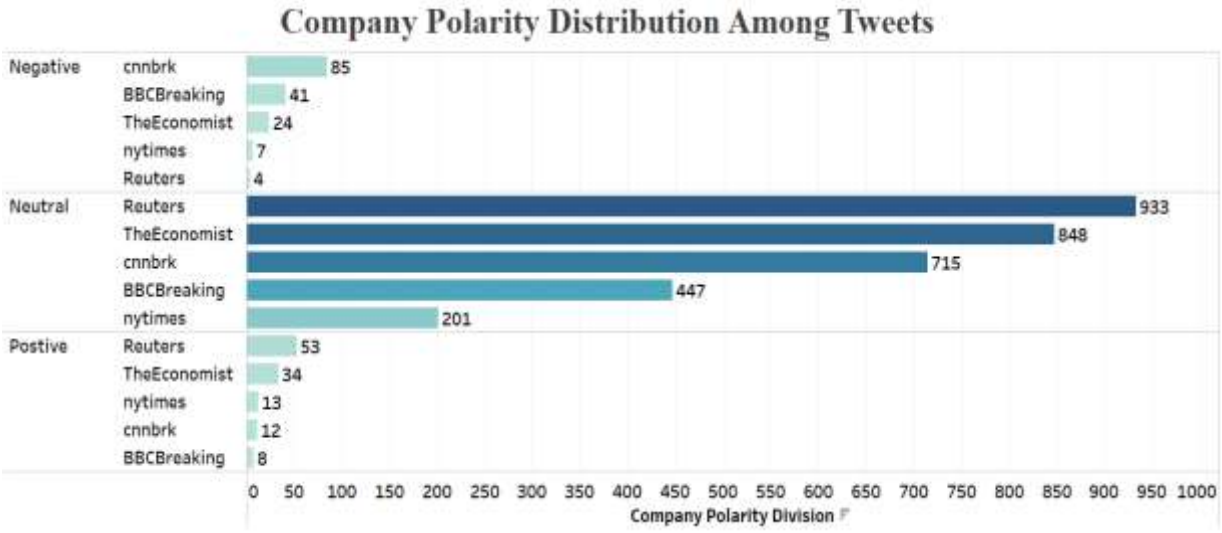


*Figure 3 Polarity Distribution among tweets categorized among companies*

# Appendix B:

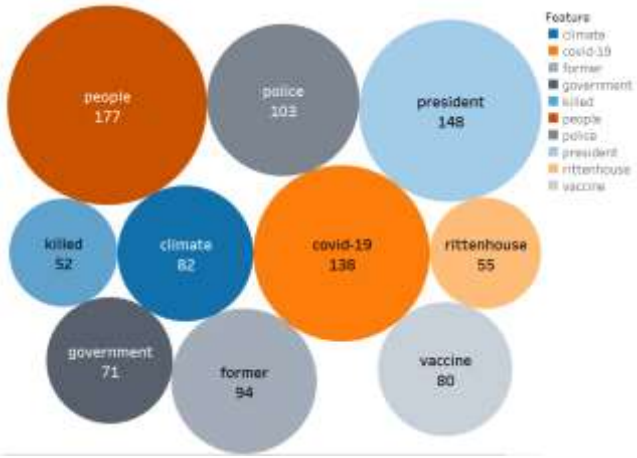


*Figure 4 Top Ten most frequent words in our document feature matrix*

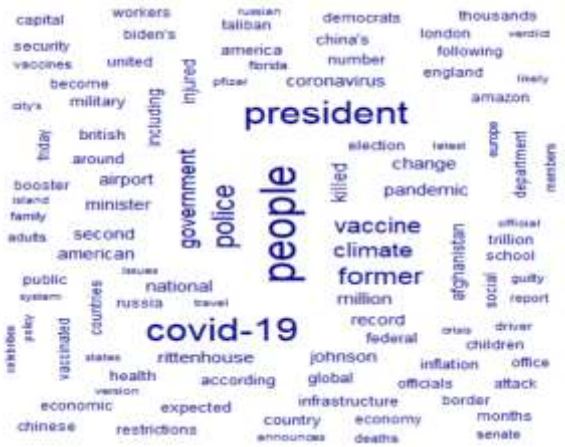


*Figure 5 Word Cloud of all tweets*

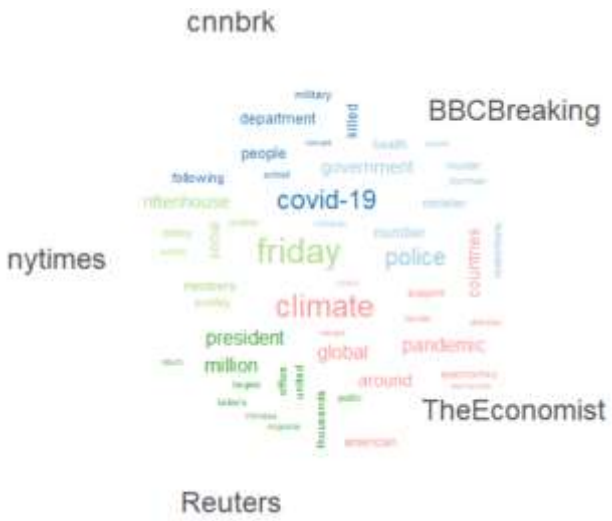


*Figure 6 Word Cloud - words categorized by company*

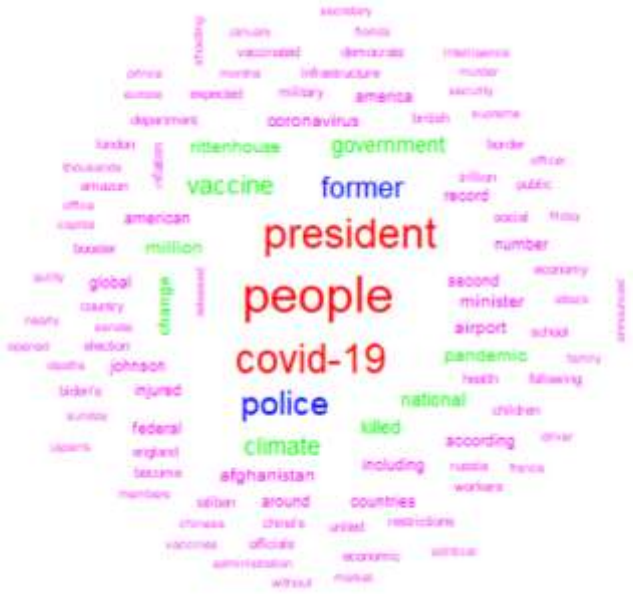


*Figure 7 Word Cloud - words color-coded by frequency in tweets*