Text Mining: UK General Election 2019

By Alistair Bernhard

Executive Summary

Currently the United Kingdom is having a general election, I scraped Twitter data (tweets) on this topic specifically so I could run sentiment analysis, and better understand the opinions of the users. It is evident that from the sentiment analysis that there was more negative polarity than positive or neutral. This is not surprising as it is a general election, and usually these events lead to a lot of opinionated statements, with most of those opinions being conflict-ridden.

Introduction

The ability to scrape data from Twitter can be helpful for many reasons, like understanding different users’ political ideologies, to knowing their favorite sports team. This information can be used to better recognize trends, and help marketers target specific groups and demographics. While this could be dangerous. If used ethically, it is a fast and accurate way to get information on any trend or topic in society. Furthermore, I pulled data from #GeneralElection19, this information could be used by a political party to learn the fears and capitalize on them for their own gains. On the other hand, it could be used by a political party to understand people’s concerns and find ways to subside them in the hopes of gathering more support for their campaign. So, the ethics is put in the hand of the individual, to use this tool in the manner they see fit.

Descriptive Statistics

In the dataset *election.txt*, there are 34,943 rows with 7 columns: identification number, date and time, tweet (text), Twitter user identification, user follower numbers, and user following numbers. The tweets that were scraped for the dataset span from December 5th, 2019 to December 8th, 2019. There are 751,559 words in the dataset, and an average of 21.5 words per tweet. Moreover, there are 26,838 unique words, meaning that this is how many words were tweeted not counting for duplicates. In *Figure 1* we can infer that the most common word was ‘with’ which was tweeted 15,453 times. Interestingly, ‘Disabled’, ‘disabilities’, ‘learning’ and ‘Sally’ all were in the top 15 of most common words, this is because of a scandal that broke at the same time I scraped the tweets. Sally-Ann Hart, a Conservative Party candidate, came under fire for her opinions on disabled people in the workplace. She said in her speech addressing her thoughts that disabled people should be paid less as “they don’t understand money.” It is evident from the dataset that these comments became a trending topic on Twitter, as people took to the platform to give their own opinions on Sally-Ann Hart and her speech.

|  |  |  |
| --- | --- | --- |
|  | **Word** | **Count** |
| 1. | with | 15,453 |
| 2. | that | 14,882 |
| 3. | should | 14,710 |
| 4. | people | 14,648 |
| 5. | her | 13,961 |
| 6. | candidate | 13,931 |
| 7. | Conservative | 13,830 |
| 8. | the | 13,810 |
| 9. | those | 13,808 |
| 10. | view | 13,631 |
| 11. | defending | 13,625 |
| 12. | learning | 13,619 |
| 13. | Disabled | 13,618 |
| 14. | Sally | 13,614 |
| 15. | disabilities | 13,611 |

*Figure 1: Table of Most Common Words and Count*

Furthermore, after researching the top two UK political parties, it was found that the Conservative Party was tweeted about more than the Labour Party. Including uppercase and lowercase letters, the Conservative Party was tweeted about 13,840 times, this accounts for 39.6% of the dataset. While the Labour Party was tweeted about 2,795 times, this accounts for 7.9% of the dataset. There may be two reasons for the Conservative Party being so prominent throughout the dataset, the first being that they are leading in the polls, thus they have more support on Twitter. But the more likely reason is the scandal of Sally-Ann Hart, as she is a member of the Conservative Party, and they were trending due to her comments.

It was important to find out how many of the users were tweeting about one of the largest topics of the UK general election, ‘Brexit’. Interestingly, ‘Brexit’ was not discussed as much as expected, as it was only tweeted about 436 times, accounting for 1.24% of the tweets in the dataset.

Method

The original dataset I pulled from Twitter did not have enough tweets to get interesting results from a sentiment analysis. Therefore, I pulled more tweets, and then ran a sentiment analysis on the tweets to find out the polarity. Then I moved the polarity results into R to find out the descriptive statistics, and make a plot showing polarity distribution.

Sentiment Analysis Results – Polarity

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in text. This is helpful as it allows us to determine whether the writer’s attitude towards a specific topic is positive, negative or neutral otherwise known as – polarity.

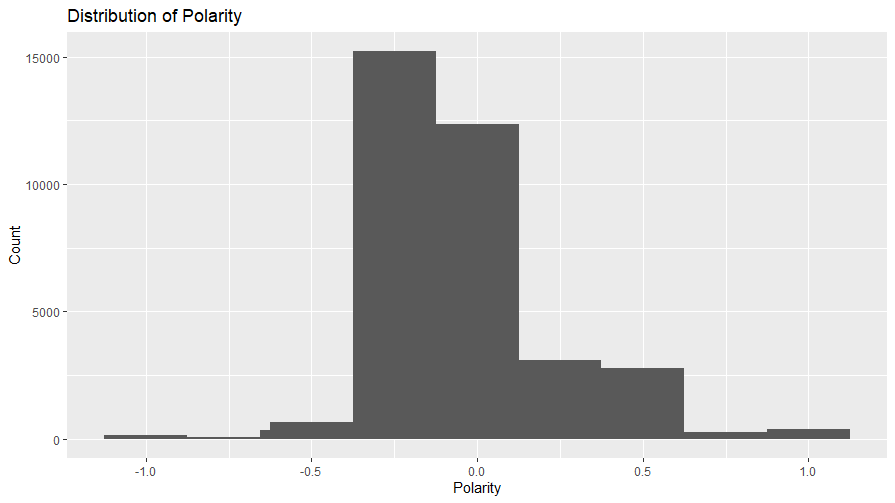
|  |  |
| --- | --- |
|  | **Negative Polarity** |
| **Minimum:** | -1.000 |
| **1st Quartile:** | -0.20000 |
| **Median:** | -0.20000 |
| **Mean:** | -0.21258 |
| **3rd Quartile:** | -0.20000 |
| **Maximum:** | -0.00119 |

*Figure 2: Negative Polarity Descriptive Statistics*

|  |  |
| --- | --- |
|  | **Positive Polarity** |
| **Minimum:** | 0.00050 |
| **1st Quartile:** | 0.16666 |
| **Median:** | 0.30000 |
| **Mean:** | 0.33122 |
| **3rd Quartile:** | 0.48750 |
| **Maximum:** | 1.00000 |

*Figure 3: Positive Polarity Descriptive Statistics*

After running some code in R on the 34,943 tweets, 16,885 were negative, 8,091 were positive and 9,967 were neutral. This is not out of the ordinary as the tweets that were scraped were surrounding the topic of the UK general election. Usually, elections in all countries can lead to divisive language, and it is not strange to read information online that is negative towards a party or candidate. This is visually evident from *Figure 4* below; the plot is close to symmetrical with a slight right skew. Most of the tweets were between -0.10 and -0.30, this is supported by *Figure 3*, as the mean was -0.21258.



*Figure 4: Distribution of Polarity*

Future Dimensions

In the future, I would have pulled more tweets, as although I have close to 35,000 tweets, it would be great to run sentiment analysis on 500,000 tweets or more. I think this would give me more interesting results, as I would have had a much larger sample size.

I will not be being using this dataset for my capstone, but I will use the skills I learnt and the code for my capstone. I will hopefully be working with a local Missoula business and I would like to scrape tweets on their hashtags and run a sentiment analysis to include it in my capstone presentation.

Appendix

Check Jupyter Notebook for python code

R code:

sent = read.csv("sentiment.csv")

ggplot(data = sent, aes(sent$polarity)) +  
 geom\_histogram() +  
 stat\_bin(bins = 9) +  
 labs(title = "Distribution of Polarity",  
 x = "Polarity",  
 y = "Count")

sent$polarity = as.numeric(sent$polarity)

neg\_pol = sent %>%

select(polarity) %>%  
 filter(polarity < 0)  
summary(neg\_pol)

pos\_pol = sent %>%  
 select(polarity) %>%  
 filter(polarity > 0)  
summary(pos\_pol)

neut\_pol = sent %>%  
 select(polarity) %>%  
 filter(polarity == 0)  
summary(neut\_pol)

count(neg\_pol)

count(pos\_pol)

count(neut\_pol)