T-Mobile vs Verizon   
 Twitter Sentiment Analysis

The Problem

For this project I compared T-Mobile and Verizon by collecting 1000 tweets for each phone carrier. I want to see which phone carrier is more popular based on twitter mentions.

The Motivation

I am interested in segment analysis, text mining area so twitter was a good place to start. Anyone can easily search a term and get mentions about the search term.

The Methods

Twitter gives limited free data through its developers account. For this project, I installed twitteR package and opened a twitter developer account. With provided keys by the twitter, I connected R program and my twitter account than I started to search my search terms. I used searchTwitter function from twitter package and set number of mentions to 1000. SearchTwitter function gives mentions as a list of 1000 elements.

The Analysis

To be able to analyze collected twitter data, I had to extract their text and save it into a variable by using sapply(tmobiletweets, function(x) x$getText()). After this step I have loaded positive and negative words list of Hu and Liu’s “opinion lexicon” categorizes nearly 6,800 words as positive or negative. The lexicon consists of two text files, one containing a list of positive words and the other containing negative words. Those two files will be attached to my projects as

"positive-words.txt"

"negative-words.txt"

* **Implementing sentiment scoring algorithm;**

To score each tweet, I have used Jeffrey Breen’sscore.sentiment() function which uses laply() to iterate through the input text. It strips punctuation and control characters from each line using R’s regular expression-powered substitution function, gsub(), and uses match() against each word list to find matches. At the end returns a data frame including mentions and their score.

Score is calculated by counting the number of occurrence of “positive” and “negative” words in a tweet. If a tweet has one positive word score will give +1, if tweet has a negative one word score will going to be -1. To calculate the score for a tweet that has more than one positive or negative word we sum the all positive and the negative scores.

After getting the results I plot histogram of both T-Mobile and Verizon sentiment score as shown in Figure 1.

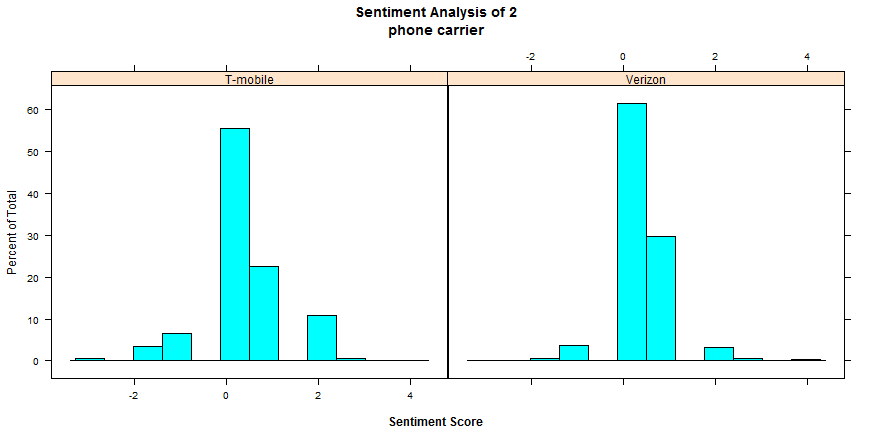


Figure 1

When we look at the histograms of both companies we will not see such a big difference. Most of the scores fall on the “0” which means neutral. That might be mentions with no positive or negative impressions or equal amount of negative and positive in the same tweets. When I went through some of the tweets I saw major neutral tweets mentioning sales or products of the company.

From the histogram we can see that Verizon has little bit higher +1 score tweets than T-Mobile whereas Verizon has little bit less +2 score tweets than T-Mobile.

To see the accuracy of the results I also plot the boxplot of the both company’s tweets sentiment scores. As expected both companies boxplots were also similar. This was a way of validation of my calculations and results. The boxplot of T-Mobile and Verizon tweet’s score is illustrated in Figure 2.

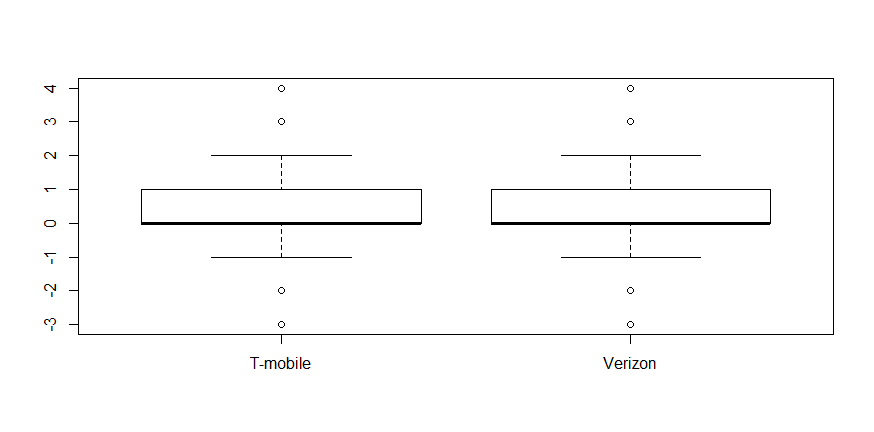


Figure 2

* **Word-clouds**

Word-clouds are useful tool that allow us to visualize most of the words associated to specific search terms in twitter. In the second part of my project I wanted to see what type of words comes when people talk about T-Mobile and Verizon. In order to accomplish this I used my collected data from twitter and extracted the text than formed a corpus.

Next, I did some cleaning and manipulation on the corpus including converting the lowercase, removing punctuations and numbers, removing URL’s and white spaces, finally cleaning from stop words.

Corpus was ready to become termdocument matrix and to be stemmed. Finally by using word-cloud package and wordcloud function I was able to plot my word-clouds for both companies.

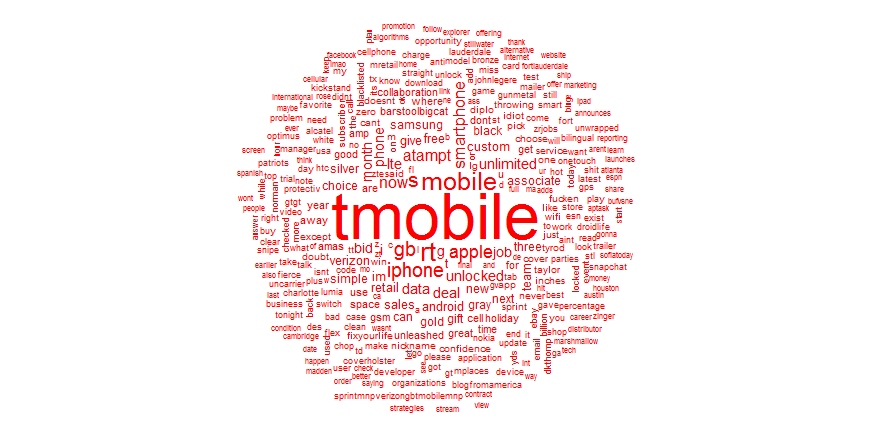


Figure 3

Word-clouds gives words that have been used most frequently. For my word-cloud I set frequency as five, so all the words that seen in the cloud is used five or more than five times. In T-Mobile word-cloud at Figure 3 above, we can see that ;apple, mobile, smartphone, gb, data words are used more frequently so their text size is relatively large. I was expecting to see some emotion words also but because of major neutral mentions I couldn’t see much.

For the Verizon word-cloud most frequent words after Verizon are; gb, iphon, black, silver, apple, unlock, deal and smartphone as pictured in figure 4.

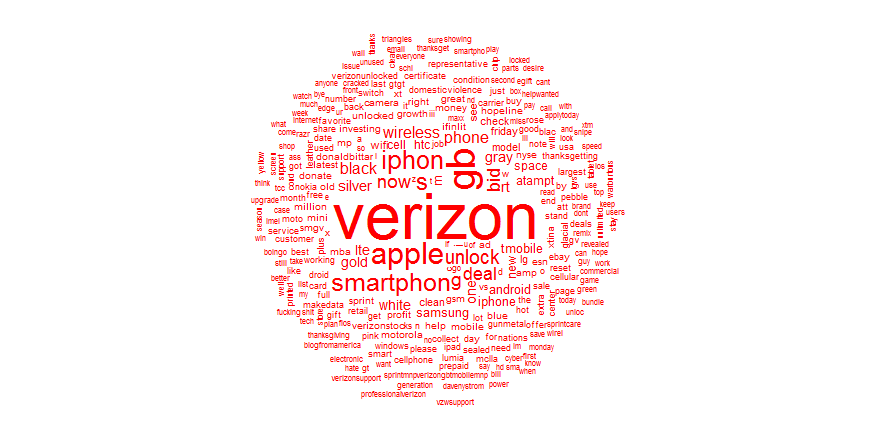


Figure 4

Conclusion

As a result, my twitter sentimental analysis has shown that both companies have similar score mentions. There is slightly different ratios on +1 score which Verizon shows higher result than T-Mobile and +2 ratio T-Mobile has higher percent of tweets than Verizon. Having large amount of neutral “0” score may indicate neutral mentions or equal amount of positive and negative words in the same tweets. Considering Black Friday season I have realized that both companies’ sale mentions are majorly falling to this neutral area.