

Text mining

Sentimental Analysis

Gautam Bhowmick

gautam.bhowmick@capgemini.com

|  |  |
| --- | --- |
| **COVER PAGE** | |
|  | |
|  | |
| Twitter   * An online social networking service that enables users to send and read short 140-character messages called “tweets" (Wikipedia). * Over 326 million monthly active users (as of Oct’2018). * Creating over 500 million tweets per day. * Percentage of Twitter users on Mobile (80%). * Number of Twitter Daily Active Users (100 million). * Twitter can handle 18 quintillion user accounts. | **“Text Mining Story”** behind the project- ***Gun violence in the United States.***  **Objectives** of our project- General sentiments of people about gun violence and how were they are reacting to this?  **Target Audience:**  Anyone who is interested in Gun politics in the United States.  **Dataset selected:**  The dataset selected for this project are the 10754 tweets pulled from twitter between Dec 02,2018 and Dec 12,2018. The data scraped from tweets were pulled into a csv file. |
|  | |

**INDEX (Table of contents)**

1. Introduction- What is Twitter Sentiment Analysis and Why are we doing this?
2. Limitations of Twitter Sentiment Analysis
3. Problem Description / Objective of the project
4. Dataset Description and Relevant variables used in the analysis
5. Overall Methodology Overview
6. Process Steps Implemented
   1. Tweet Extraction
   2. Text Cleaning
   3. Building Term Document Matrix- finding most frequent words
   4. Word Associations
7. Visualizations of text data- Networks and Word Clouds
8. Sentiment Analysis
   1. Bag-of-Words approach of tweet words classification
   2. Sentimental scoring based on Bag-of-Words approach on whole tweets
9. Sentiment polarity trend over time
10. Conclusion
11. R Packages Used
12. Online References
13. Book References

**Introduction-What is Twitter Sentiment Analysis and Why are we doing this?**

Twitter has been a very useful public platform to mine people’s opinions all over the world. Twitter Sentiment Analysis is the process of determining the emotional tone behind a series of words and is used to gain an understanding of the attitudes, opinions and emotions expressed within an online mention.

The applications of sentiment analysis are endless. It is extremely valuable when it comes to business analytics, especially marketing sectors where understanding people’s pulse plays a vital role in getting the edge over competitors. Thousands of text documents can be processed for sentiments in seconds, compared to the hours it might otherwise take a team to manually complete the same, because it is so efficient and accurate, many businesses are adopting text and sentiment analysis and incorporating it into their processes.

**Limitations of Twitter Sentiment Analysis:**

Having laid out the benefits and purpose, it also becomes mandatory to point to the limitations-

* The Twitter Search API can get tweets up to a maximum of 7 days old.
* Not effective in detecting sarcasm.
* Cannot get 100% efficiency in analyzing sentiment of tweets.
* Can only retrieve a maximum of 1000 tweets per query without authenticating via OAuth before receiving a 403 error or timeout.
* Giving a hash tag under the wrong category will still give results: No error message!
* Anaphora Resolution: The problem of resolving what a pronoun, or a noun phrase refers to. "We watched the movie and went to dinner; it was awful." What does "It" refer to?
* Parsing: What is the subject and object of the sentence, which one does the verb and/or adjective refer to?

**Problem Description / Objective of the project:**

**Gun violence** in the **United States** results in tens of thousands of deaths and injuries annually. In 2015, [over 13,000 people](http://www.bbc.com/news/world-us-canada-34996604) in the US were killed by firearms — but an even more staggering number is those who have survived shootings: 26,819.

2018 is worst year on record for gun violence in schools.

**Objectives:**

* General sentiments of people about gun violence and how were they are reacting to this?
* How do people, in general, feel about gun violence in united states?

**Dataset Description and Relevant variables used in the analysis:**

The dataset used for this project was the **10,754** **tweets** pulled from twitter. **Retweets were filtered out** during the initial pull to avoid repetitions which could cause skew. Twitter allows pulling only 7 days of data hence, the tweets were carefully selected and pulled between **Dec 02,2018 and December 12,2018**. The data scraped from tweets were pulled into a csv file.

The variables pulled with the **10,754** tweets were-

* Tweet text: The original words of the tweet
* Favorited: Boolean (True/False) value saying if the tweet was liked or not.
* FavoriteCount: Number of times a tweet was favorited.
* Created: Date of the tweet (datetime format).
* Truncated: Boolean (True/False) value showing if tweet was truncated.
* Retweeted: Boolean (True/ False) but always False since we filtered out retweets at the beginning.

Other variables include- ReplyToSN, ReplytoSID, ID, ReplytoUID, StatusSource, ScreenName, retweetCount, longitude, latitude. Among these, the relevant variables used in the analysis were just 2-

1. Tweet text – for Sentiment Analysis of tweet words as well as to determine polarity of tweets as a whole.
2. Created- for Sentiment trend over time in the 7 days.

*Over 1.15 M people have been killed by guns in the USA since John Lennon was shot and killed on Dec 8, 1980. That’s the reason you may see a spike between 8 Dec to 9 Dec.*

|  |
| --- |
| **Tweets over time** |
|  |

|  |
| --- |
| **Count of the number of tweets by day, hour** |
|  |
| **Word Cloud (Raw Tweets)** |
|  |

**Methodology Overview**

1. Extracting tweets from the Twitter website with R and the ‘twitteR’ package.
2. With the ‘tm’ package, cleaning text by removing punctuations, numbers, hyperlinks and stop words, followed by stemming and stem completion.
3. Build a term-document matrix and analyzed the most frequently used words.
4. Took 2 approaches to analyze sentiments:
   1. Unsupervised Bag-of-words approach by comparing to pre-loaded word database and assigning sentiment score- positive, negative or neutral and calculating overall score.
   2. Analyzed sentimental score on whole tweet using Bag-of-words approach.
5. Using ‘wordcloud’ package, built positive, negative as well as overall word clouds.
6. Network analysis was also performed to visualize the tweet sentiments in general.

**Process Steps Implemented**

1. **Tweet Extraction**

(i) Create twitter application

(ii) twitteR package- Provides an interface to the Twitter web API

(iii) ROAuth package- R Interface For OAuth

(iv) Create twitter authenticated credential object (using key from step (ii) and cacert.pem certificate): It is done using consumer key, consumer secret, access token, access secret.

1. **Text Cleaning**

The tweets are cleaned in R by removing:

* Extra punctuation
* Stop words (Most commonly used words in a language like *the*, *is*, *at*,
* *which*, and *on*.)
* Redundant Blank spaces
* Emoticons
* Stemming is also done wherein words like giving and give are stemmed or truncated and then Stem Completion was also performed to produce a meaningful word

1. **Building Term Document Matrix- finding most frequent words**

The main structure for managing **documents** in tm is a so-called Corpus, representing a collection of text **documents**. A corpus is an abstract concept, and there can exist several implementations in parallel. Creating a DTM (or TDM) in **R** is as simple as: dtm <-**DocumentTermMatrix**(docs). This creates a **term document matrix** from the corpus and stores the result in the variable dtm. One can get summary information on the **matrix** by typing the variable name in the console and hitting return: dtm.

The top 10 most common words found using TDM were-

|  |  |
| --- | --- |
|  |  |

1. **Word Associations**

Word associations of interesting words- in this case- ‘kill’ and ‘gun’ were found using FindAssoc function.

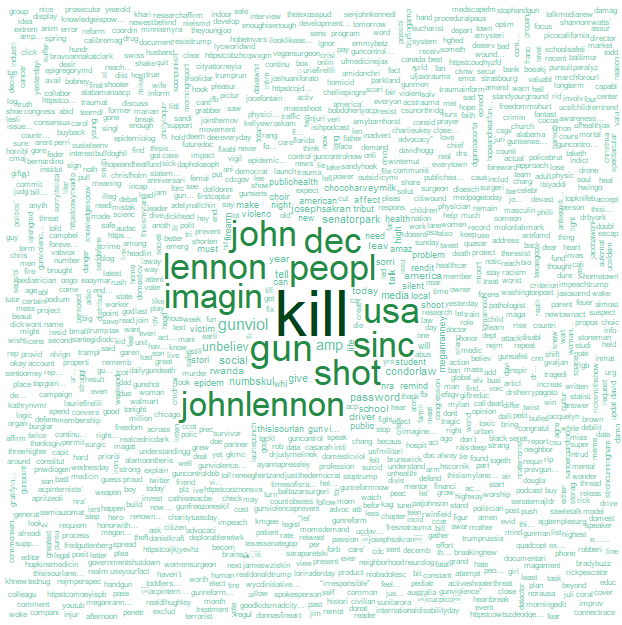
Interestingly, the word ‘kill’ an d ‘’gun’ seems to have been associated mostly with ‘shot’, ‘gun’, ‘peopl’ and ‘usa’. We could consider this as a negative pulse of people towards the gun violence.

It’s hard to infer anything, but more on this discussed in the Sentiment Analysis and Conclusion sections!

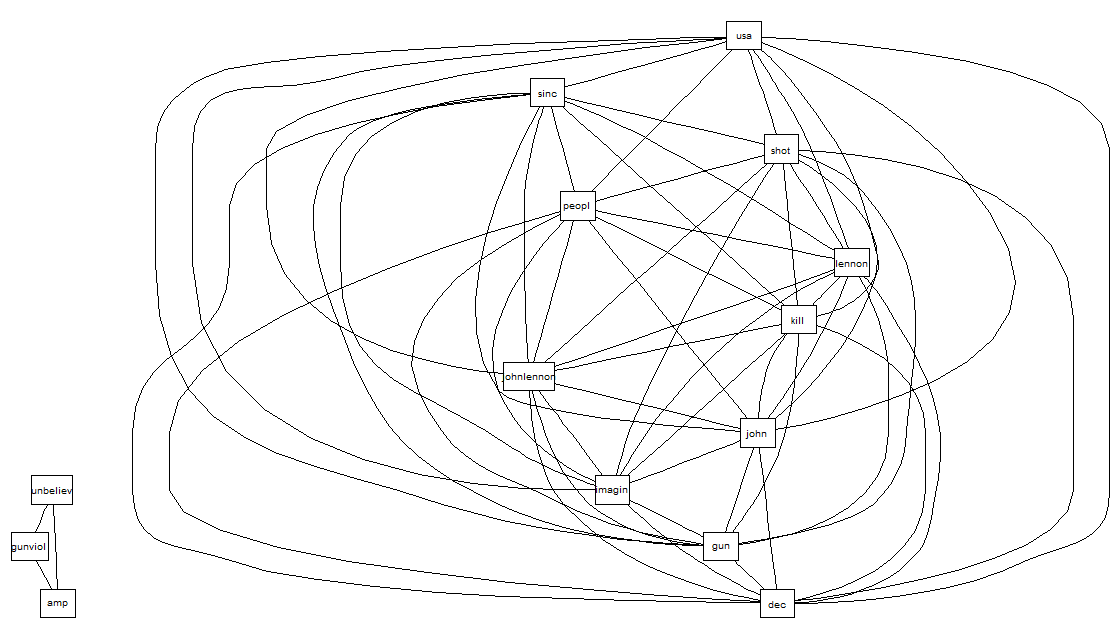
|  |  |
| --- | --- |
|  |  |

**Visualizations of text data- Networks and Word Clouds**

A Word Cloud of the most frequent words used on tweets between these 7 days on this topic showed us-



We see that the wordcloud of most frequent words mostly shows negative words like- ‘kill’,’gun’,’shot’ etc. Which would explain the overall higher negative score that we are going to see in Sentiment Analysis later.



graphviz library was used for plotting R graph objects from the graph package. I didn’t use weight in graph object since occurrences of words are high. Network diagram shows strong connectivity among ‘Kill’, ‘gun’, ‘shot’, ‘peopl’ and ‘usa’.

**Sentiment Analysis**

Considering the limitations of twitter sentiment analysis discussed prior, I took 2 steps procedure to validate authenticity as well as to make the most accurate conclusion on this analysis.

1. **Bag-of-Words approach of tweet words classification**

Using the bag of words approach, we uniquely identify each word (mostly adjectives or adverbs) as either belonging to positive word list, or negative word list.

1. List of positive words: <http://ptrckprry.com/course/ssd/data/positive-words.txt>
2. List of negative words: <http://ptrckprry.com/course/ssd/data/negative-words.txt>

A couple of additional positive and negative words were added to suit the context.

After the tweet text has been segmented into sentences, each sentence has been segmented into words, and the words have been tokenized and normalized, and made a simple bag-of-words model of the text. In this bag-of-words representation considered individual words into account and give each word a specific subjectivity score.

I was curious on what words were most commonly used for the **positive tweets** and what words were most commonly used for the **negative tweets**, just to get a feel of the pulse as well as to test the accuracy of our sentiment analysis on a high-level.

|  |  |
| --- | --- |
| **Positive Word Cloud** | **Negative Word Cloud** |
|  |  |
|  | |

|  |
| --- |
|  |

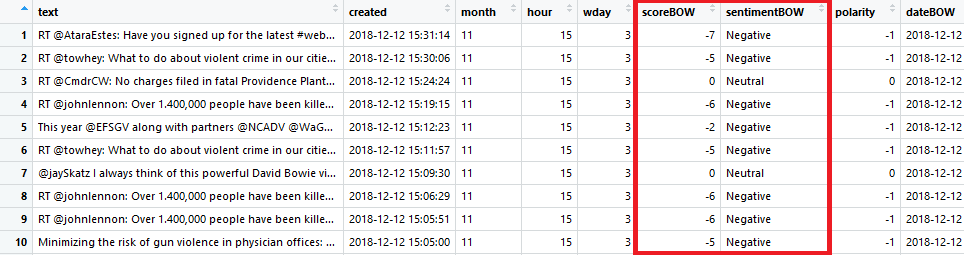
Taking a closer look, negative word cloud predominantly does show only negative words, like- ‘kill’, ‘shot’, ‘gun’ etc. and so does the positibe word cloud with words like ‘silent’, ‘thank’ etc.

We also have to consider the limitations of Sentiment Analysis which is discussed later.

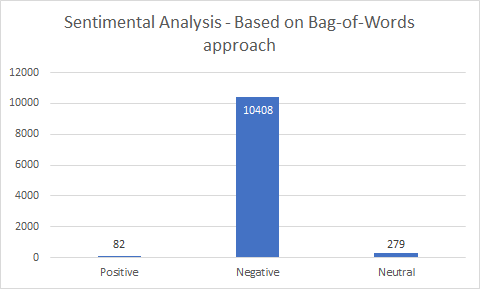
1. **Sentimental scoring based on Bag-of-Words approach on whole tweets**

Based on subjectivity score given to bag-of-words representation now it can be looked up in a sentiment lexicon. If the total score is negative the text will be classified as negative and if its positive the text will be classified as positive.

A snapshot of the tweet vs. scoring dataset



The resulting tweet word sentiment classification is plotted as below

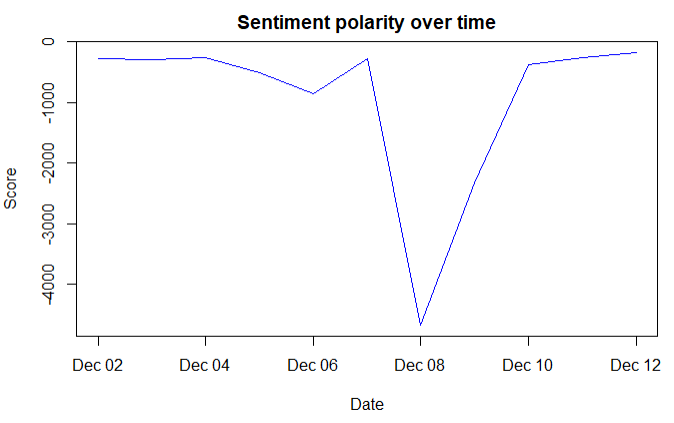


Negative words in the tweets seem to be predominating followed by neutral words, which seem to have a higher distribution than the positive words.

We can start feeling pretty good about ourselves because both outputs for most positive and most negative sounds about right!! Positive and negative wordcloud were plotted above and which are discussed in the *Bag-of-Words approach of tweet words classification* section of this report.

**Sentiment Polarity Trend over Time**

So now, the question arose as to- we now have the overall distributions of the positive, negative and neutral sentiments and have a wordcloud showing the most frequent positive words and most frequent negative words used. So, we now have an idea as to the pulse of our general audience, but wait- is that a standardized conclusion? First step-we score the sentiments and plot them. As is easily inferable from this plot, we see a lot of more negative scores than positive scores. Then we get into pulling in the date variable from each of the tweet and plot a trend on time graph as shown. As is evident, to confirm our doubts, though the other 2 approaches and sentiment scores in this approach had learned more towards classifying people’s predominant emotions towards the #**gunviolence** as negative, we see that that has not always been the case! There was a marked few positive trend and again almost immediately moved to negative drop in scores around 8th, 9th and 10th Dec *since John Lennon was shot and killed on Dec 8, 1980 and over 1.15 M people have been killed by guns in the USA since he was dead.*



Again, this is a short span of time and further extended analysis could give us more clarity on this topic. We cannot be 100% accurate, but, nevertheless, we seem to have achieved our goal of detecting the negative pulse of people towards the gun violence.

This trend graph nails the conclusion!

**Conclusion**

Accuracy of Sentiment Analysis is compromised due to its limitations. Hence, though we see a higher negative sentiment than positive sentiment (disregarding the neutral ones) in Bag-of-Word sentimental analysis approach, the trend graph nails the exact conclusion on how people feel and felt about the #**gunviolence**, in general. There was positive sentiment, but it didn’t actually mean that people were feeling positive or would continue to feel positive about it. Further analysis and more tweets would be able to give a clearer conclusion on this topic. But for starters, I think we are on the right path now!

**Packages Used**

● **openssl:**Toolkit for Encryption, Signatures and Certificates Based on OpenSSL

● **xts:** eXtensible Time Series

● **httpuv:** HTTP and WebSocket Server Library

● **SnowballC:** Snowball stemmers based on the C libstemmer UTF-8 library

● **Rgraphviz:** Provides plotting capabilities for R graph objects

● **topicmodels:** Topic Models

● **twitteR :** Provides an interface to the Twitter web API

● **stringr** : String operations in R

● **ROAuth :** Provides an interface to the OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice.

● **RCurl :** Provides functions to allow one to compose general HTTP requests and provides convenient functions to fetch URIs, get & post forms, etc. and process the results returned by the Web server.

● **ggplot2 :** An implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics: conditioning and shared axes are handled automatically, and you can still build up a plot step by step from multiple data sources.

**● tm :** A framework for text mining applications within R.

● **wordcloud :** visual representation in the form of wordcloud where size of the word is proportional to the frequency of words used in the tweets

● **plyr :** Tools for Splitting, Applying and Combining Data

● **dplyr:** dplyr is a grammar of data manipulation

**References:**

<https://twitter.com/hashtag/gunviolence?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Ehashtag>

<http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know>

<https://www.omnicoreagency.com/twitter-statistics/>

<https://en.wikipedia.org/wiki/Gun_violence_in_the_United_States>

<https://rstudio-pubs-static.s3.amazonaws.com/90345_36e829555b464fb08ed888d1678827f9.html>

<https://rpubs.com/williamsurles/316682>

<https://rdatamining.wordpress.com/2012/05/17/an-example-of-social-network-analysis-with-r-using-package-igraph/>

<https://www.r-bloggers.com/visualising-twitter-user-timeline-activity-in-r/>

<http://blog.kaggle.com/2017/10/05/data-science-101-sentiment-analysis-in-r-tutorial/>

**Book References:**

Practical text analytics: Interpreting text and unstructured data for business intelligence. Kogan Page.

Jockers, M. (2014). Text analysis with R for students of literature. Springer.