# **Sentiment Analysis on USA’s 45th President.**

# 

|  |  |
| --- | --- |
| **Name** | **CWID** |
| Abhishek Dutta | A20379160 |

**Type of Your Projects:** Data Analysis/Mining, Application Implement/Development, Research Project

1. **Introduction**

In today’s time where the penetration of the internet has reached to such a great extent and the number of users on social media is increasing at a fast rate it has become a fashion of uploading ideas, opinions, statements, emotions and various other forms of textual content over various social media sites. Many smart companies have therefore started utilizing this data over the internet to extract something more out of it that would in turn help their firms to analyze things related to their business and predict about future strategies simultaneously.

Micro blogging sites is a platform where people post textual data such as their opinions on various important issues, contents, expressions, moods, analysis of products their interpretation of things and reviews.

Considering, the best and most famous micro blogging site in particular –Twitter. Let’s see what all things can it be used for:

1)Twitter is used by a lot of people these days and therefore we get a huge number of opinions coming from people with different background that could be used for analysis of various subjects.

2)Its audience varies from normal users to celebrities, politicians, sports personalities, media people business men and other such people whose views, opinions and ideas really count a lot and have a huge impact globally.

Therefore, we have taken up the task to perform a sentimental analysis on such a personality who happens to be the President of one of the most influential country, the United States of America. The main reason behind taking up this topic is that Mr. Trump being the president of America holds a very crucial position and after 2016 election, people around the world have started following news related to him very closely and started sharing their views on him frequently. People from different parts of the world have opinions and views on each decision that he makes or bill that his team proposes. Therefore, it gives us a great opportunity to analyze what people think about him and about the decision that he makes. Based on the tweets, we are planning to extract the sentiments of the people regarding him, his decisions and what better topic can it be to analyze than our very own President Mr. Trump 😊

1. **Data Sets**

We are planning to extract data from Twitter using **#President Trump**. For this we would have a data set of 5000 entries out of which 80 % would be falling under training and 20% test. As of now, we are trying different ways to extract the output and will use the most efficient and accurate way for our project. One way to download the tweets from twitter is by sitting up and API on twitter.com based on few keys.

1. **Research Problems**

**Sentimental Analysis**- It is basically the detection of attitude, behavior, sentiment or emotions conveyed in the form of textual data. It can be considered as a branch of NLP (Natural language processing).

* Now since our datasets would contain texts, emoticons, urban lingos and text in different language therefore it becomes very important to differentiate and filter out not the important words for sentimental analysis. Since tweets are messy, cleaning tweets are very important.
* Once we have cleaned the texts by removing the irrelevant data then we would have to figure out whether the left words fall under positive or negative category and then how to come up with whether the tweets fall under positive or negative sentiment as some tweets may have both positive and negative words in them.
* Since the data set we download is unlabeled therefore we would require clustering technique to perform various tasks such as Clustering tweets with similar contents and sentiments and others.

1. **Potential Solutions**

* Once we download the tweets it’s very important to clean it therefore for doing that we are planning to use R packages from CRAN like stringr and its functions. Once we clean the tweets our tweets data would be just vector of character strings.
* Once we gather the tweets, to perform sentiment analysis, we would see the text and create a sentiment score which carries the positive words score (+1) and negative words score (-1) in a tweet. The words in the tweets that don’t fall under negative or positive word list would have a score 0. These words are of no use to us in calculating the score or analyzing the sentiment. Now in a tweet, If the total positive score is greater than the total negative score then that tweet would fall under positive sentiment and vice versa.
* We would be using the AFINN Library for the sentimental analysis which was researched by Arup Nielson (2015) It gives a word scores from +5 to -5 towards sentimental analysis or short text.

<https://finnaarupnielsen.wordpress.com/2011/03/16/afinn-a-new-word-list-for-sentiment-analysis/>

* The data mining technique used to group the positive and negative words is Clustering technique. We use clustering as the tweets that we are extracting will not be having structures or labels. Here we will discover underlying structures from the unlabeled data. Clustering tweets with similar contents and sentiments and others.

**Example:**

-> I **don’t** like Trump is **bad** as a president and is **injurious** for the society.

->Trump would create **problems** for immigrants.

->Trump is proving to be a **Good** president.

Here, we can clearly see from above tweets are having bad words and good words therefore would be clustered as a negative sentiment and Positive Sentiment with scaling and will fall in respective bucket.

Like the Words like don’t, bad, not fall in scale 1 Negative words and hence be grouped together. Whereas the words like problem has more impact and similar words will be grouped together.

How the process goes:

**Text Mining -> Setting up a Twitter API/ any other means to download twitter data -> Summarize the traits of the tweets -> Cleaning the tweets -> Storing the text data -> Clustering tweets -> Sentimental Analysis is performed -> required outcome is obtained.**

As of now we are planning to use R programming tool to perform the above task and with the help of packages from CRAN we would be able to perform the tasks related to extracting data from twitter, text cleaning, summarizing, Word cloud, Sentiment analysis and others using textual data in R. Belo mentioned are the name of few of those packages :

**twitteR, stringr, Word cloud, sentiment, igraph and others.**

**Few of the secondary packages would come along with them as well.**

1. **Evaluations**

Sentimental Analysis will be used to find out the emotion conveyed via the “tweets”. Here we will be using the Sentimental Score to find out if the word is a positive or negative. Depending upon this we will find about the whole sentence like if this will have more negative or a more positive content in it. The Sentimental score will give if the Sentence structure will have the data as positive or negative based on the Maximum weight. Here the we would be using a function that search through the text in the tweet and look for the words that matches whichever word list (positive or Negative) we want to use. The Function will return original sentence with their scores.

1. **Expected Outcomes**

Based on the sentimental analysis we perform, we can do a proper analysis on how the regular public, celebrities, politicians, global leaders and other influential people react to Mr. Trump’s decisions, proposals, tweets, views on global issues and other matters. This will also give us an idea about how many people support him on his decisions and how many people don’t. Expected outcome is to have a positive feedback or a negative feedback about the performance of the President.