Twitter Sentiment analysis on ‘Obamacare’

-DA5020

Satish Reddy Chirra
Candidate for MS in Data Science
College of Computer & Information Science
Spring 2017

Introduction

Sentiment analysis is the task of identifying whether the opinion expressed in a text is positive or negative in general, or about a given topic. Nowadays, when the micro-blogging platforms, such as Twitter, are commonly used, the task of sentiment analysis becomes even more interesting. Micro- blogging introduces a new way of communication, where people are forced to use short texts to deliver their messages, hence containing new acronyms, abbreviations, and grammatical mistakes that were generated intentionally. Although there are several known tasks related to sentiment analysis, in this project we will focus on the sentiment analysis of '#Obamacare' using tidy text lexicons. In other words, the texts that we deal with in this project, will express a sentiment score either positive, negative or neutral sentiment.

In this project, we will use two sets of tweets: 1) Tweets collected without any specific Geocodes and 2) Tweets collected with the specific Geocodes that covers all the states of USA. The data collected from Twitter are short messages (i.e., bounded by 140 characters), an annotated topic, and the sentiment that is expressed toward it. The annotation of the topic and the sentiment was done by human annotators. I have used tidy text lexicons to get the sentiment of the tweets, that each tweet would be having a score either positive, negative or zero based on the emotions expressed in the tweet.

Motivation

Considering the current political dilemma on healthcare system in USA, knowing what are the emotions of common people on the same encouraged me to choose this project. Twitter being a most active platform in communications with a half-billion tweets sent every day, it has become an important platform for sharing news, ideas, and opinions. As the number of users increasing, microblogging platforms are becoming a place to find strong viewpoints and sentiment.

People use twitter to predict a lot of different areas. Twitter is really useful for predicting emotions of the people on the current trending topics. It is one important reason why Twitter is chosen to predict how people think about the popularity of different issues. Another reason is because Twitter serves as a worthy platform for sentiment analysis due to its large user base from a variety of social and cultural regions worldwide. This can be easily collected through its APIs (Application Program Interface), which makes it easy to build a great data set.

Data Collection

Data is collected through twitter API. As Twitter allows the users to extract only the last one week data, it helped in getting the current emotions of the people. However, there was a great deal of challenge in extracting the large number of tweets as it allows only 15 connections in a span of 15 minutes.

In this project, two kinds of datasets have been used.

1) One without using Geocodes. A data size of 15000 tweets without Geocodes has been extracted using API which includes both popular and latest tweets which can be extracted using setting the type to 'mixed'.
2) Other with the Geocodes. For the data with Geocodes, location details like latitude and longitude are collected through an online source [1], due to very high number of geocodes I have taken the mean of all latitudes and longitudes with respect to each state and defined a radius of 1000 miles for each entry to extract the tweets from all the states in USA. A data size of 908 tweets have been extracted for fifty states. As the initial tweets are not specific to any location, tweets with geocodes helps in getting the people emotions specific to different state in USA.

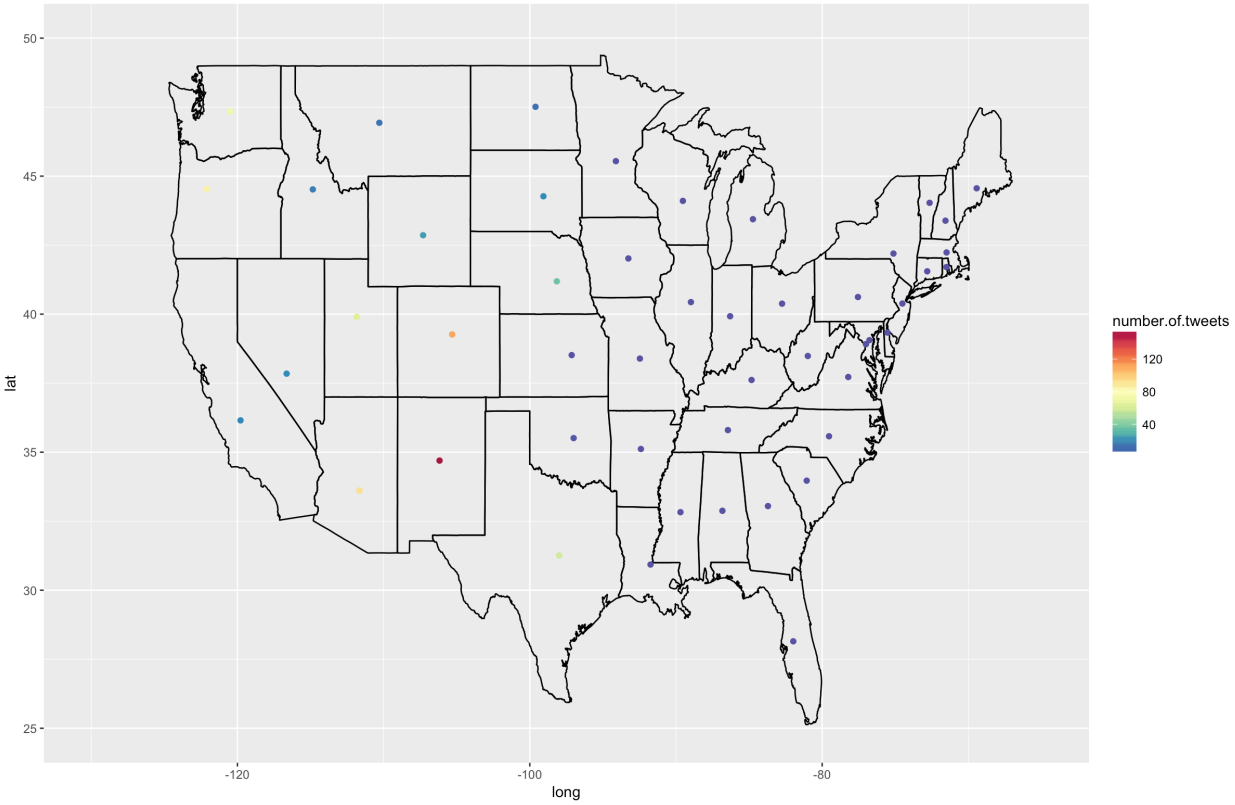

Figure 1. Map showing number of tweets collected in different state of USA on a scale of 1 to 157.

Data Cleaning

As the tweets extracted as very unstructured, it was great deal to clean the data to extract only the sensible data for the sentiment analysis. Package 'tm' has been used for the data cleaning. Below is the process followed to clean the data.

- Removing all the special characters from the tweets
- Removing the usernames and hyperlinks.
- Creating the corpus of text using all the tweets
- Removing stop words, punctuation, numbers, white spaces and other specific words.
- Transforming the tweets to lower case.

Even after the above steps, data was not in the required level of cleanliness. It was a great deal of challenge as there are two data sets and getting into the required format was tedious.

Data Exploration

As the data is in the required format, selecting a method to analyze was very difficult. Initial analysis was done to get the sentiment of the tweets without using the tidy text lexicons. In this method, defining the positive and negative words are great deal of challenge. I have used '**word vectors**' to get the cluster of words with the same sentiment. In order to get the cluster of words with similar meaning, a model of vector of 300 dimensions has been created which is trained on the entire corpus of tweets and similar words with positive and negative annotations have been found with the help of the model.

Once the words with the positive and negative annotations are found, I have used a function to get the score of the tweets which has a score range of 1:4. In search of a better approach to assign a better emotion I have found tidy text appealing, which has better score range for negative and positive scores.

Data Storing

Selecting database to store the data was relatively easy as the data size is not huge I have selected to go with **RSQLite**. A database is created and two different data sets which have been cleaned are loaded to distinct entities of the database. However, for the majority of the analysis part, it was the R environment objects which are used.

Data Analysis

Initially, once the data is cleaned I have extracted a list of words and generated a word cloud for both the data sets. I have used 'wordcloud' package to generate the same. Below is the word cloud generated from the entire word list with the total words of 200 and each have a minimum frequency of 50.

<u>Word Cloud for tweets without Geocodes:</u>

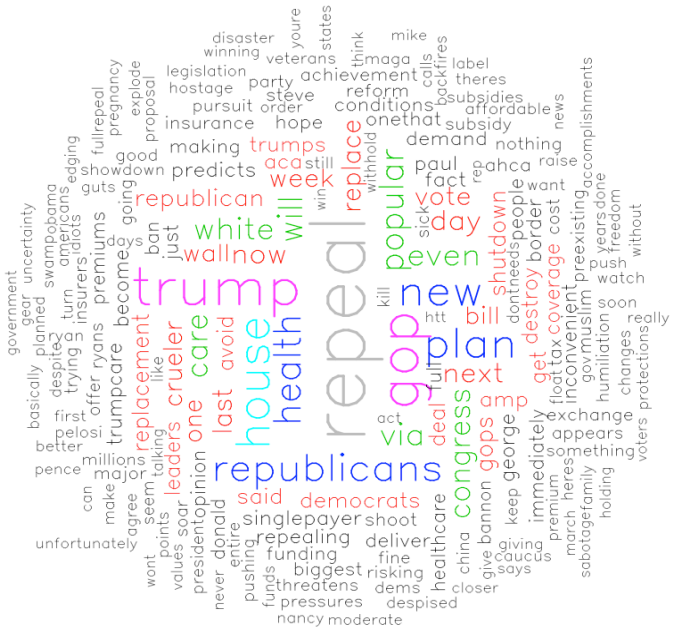

Figure 2. word cloud generated from the entire word list with the total words of 200 and each have a minimum frequency of 50 for tweets without Geocodes

Word Cloud for tweets with Geocodes:

Figure 3. Word Cloud with positive and negative words for tweets with Geocodes

After the initial analysis done on the various methods available to analyze the data it made more sense to use the tidy text for the analysis. There are three types of lexicons in tidy text package based on the Unigram,

- `AFINN` from Finn Årup Nielsen,
- `bing` from Bing Liu and collaborators, and
- `nrc` from Saif Mohammad and Peter Turney

These lexicons contain many English words and the words are assigned scores for positive/negative sentiment, and also possibly emotions like joy, anger, sadness, and so forth. The nrc lexicon categorizes words in a binary fashion ("yes"/ "no") into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The bing lexicon categorizes words in a binary fashion into positive and negative categories. The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.

I have used 'nrc' lexicons to get the number of words with the different emotions. Below is the table for the same.

sentiment <chr>	**n** <int>
negative	13208
positive	12549
trust	9667
anticipation	9219
surprise	6535
sadness	6077
fear	5991
anger	5983
joy	5889

Table 1. Number of words with different emotions from tweets without Geocodes

sentiment <chr>	n <int>
negative	3841
positive	2870
trust	1870
fear	1694
anger	1522
sadness	1465
disgust	1350
anticipation	1258
joy	988

Table 2. Number of words with different emotions from tweets with Geocodes

I have used 'bing' lexicons to get the words which contributed most to the sentiment analysis of the tweets.

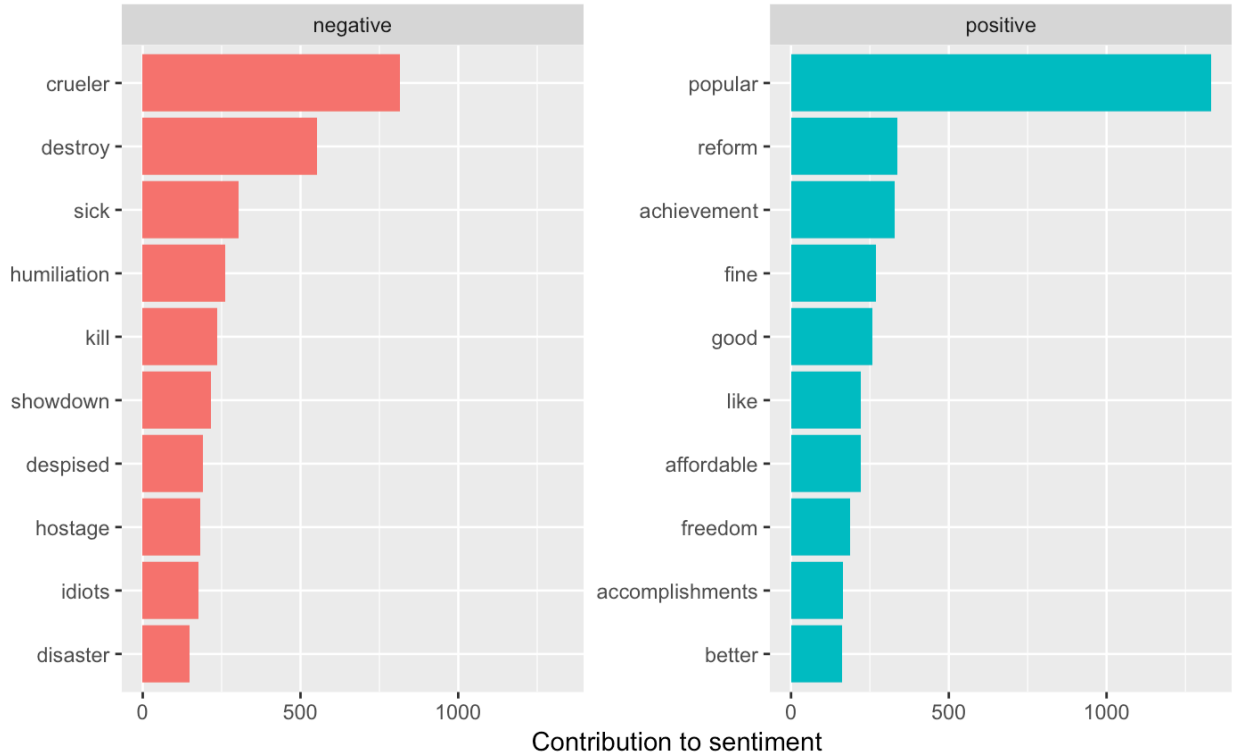

Figure 4. Words which contributed the most in overall sentiment of the analysis for without Geocodes

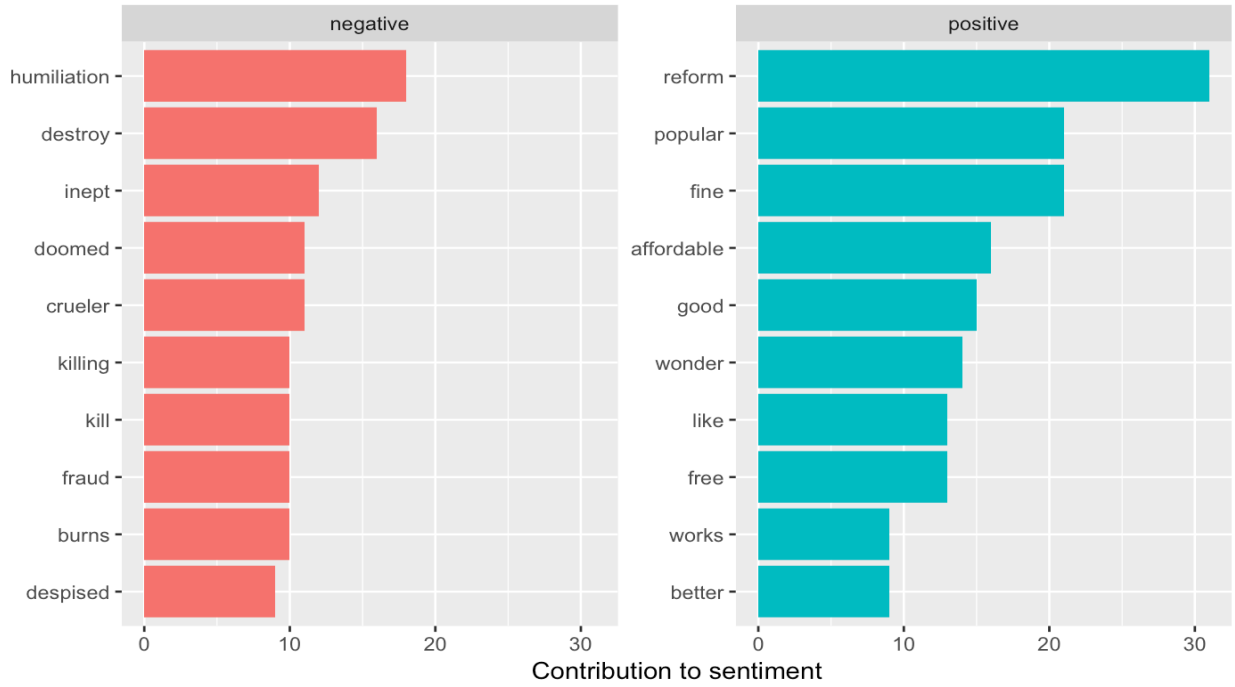

Figure 5. Words which contributed the most in overall sentiment of the analysis for with Geocodes

To get the sentiment of the tweet I have used AFINN lexicon which score for each word between -5 and 5. I have divided each tweet into words and left joined with the lexicon to the scores. Once the scores are assigned to the words I have grouped based on the tweet ID and took a sum of score which gave the total sentiment of the tweet. Below is the ggplot which indicates the sentiment of the tweets.

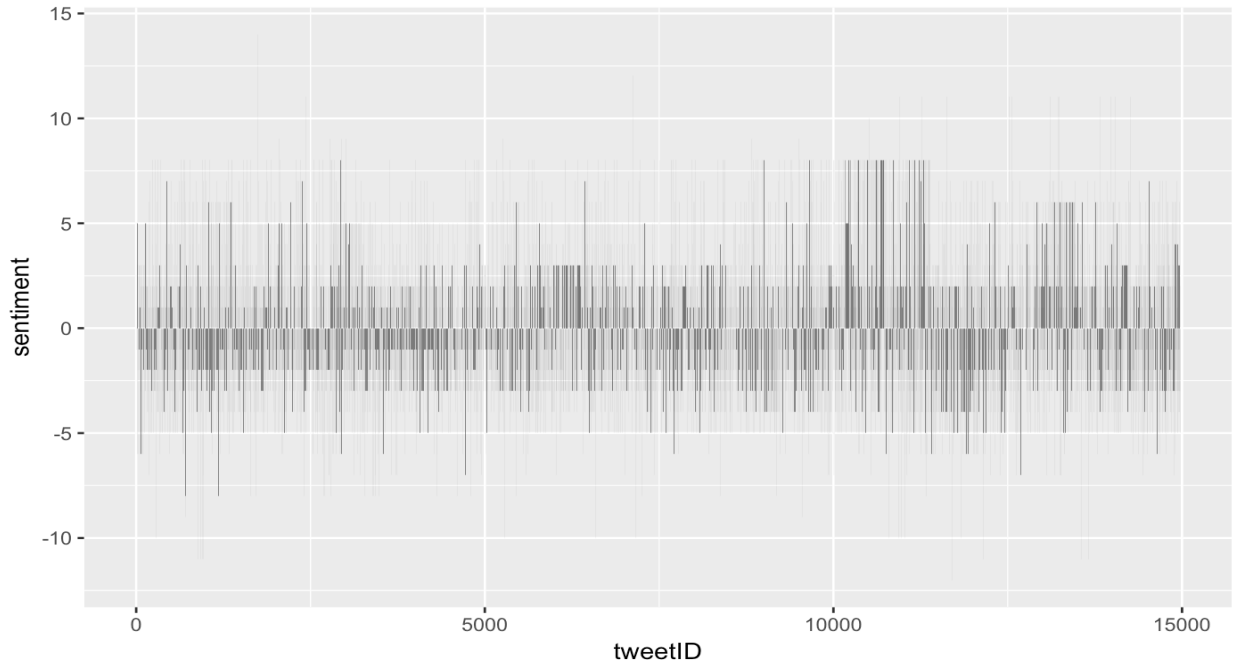

Figure 6. Sentiment of tweets without Geocodes

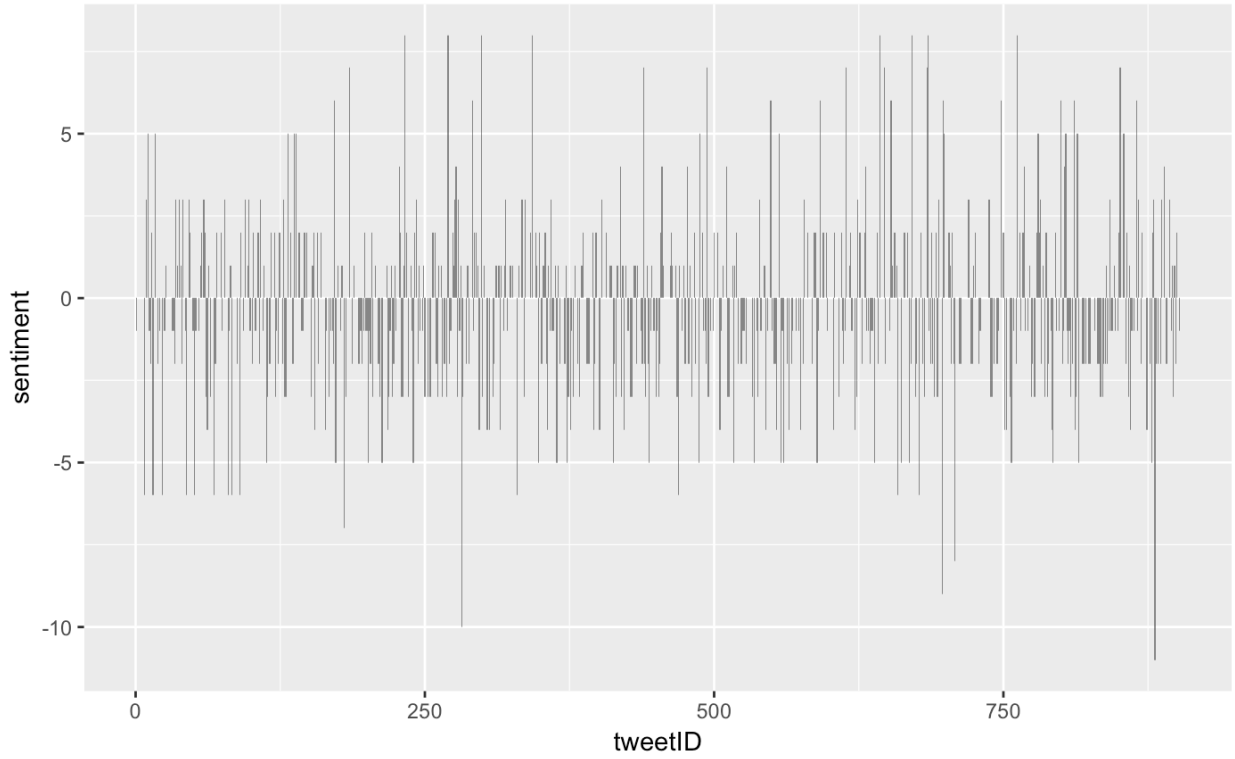

Figure 7. Overall Sentiment of the tweets with Geocodes

Future Analysis

For the future analysis, I would like to try those 300 dimensional vectors as features and give it as an input to various machine leaning algorithms especially Neural Network, to get a deeper understanding on the sentiment behind the tweets.

References

[1] GeoCodes: http://www.farinspace.com/us-cities-and-state-sql-dump/
[2] Tidytext http://tidytextmining.com/index.html