TWITTER SENTIMENT ANALYSIS

PROJECT DESCRIPTION

Sentiment analysis is a relatively new area, which deals with extracting user opinion automatically. An example of a positive sentiment is, “natural language processing is fun” alternatively, a negative sentiment is “it’s a horrible day, I am not going outside”. In this project, we propose to perform sentiment analysis on the domain of Twitter updates because of the challenging nature of this category of texts. Positive comments tend to be straightforward but negative and critical opinions tend to be expressed through sarcastic and sometimes ambiguous language. The main advantage of studying Twitter updates is that after annotating for the positive or negative sentiment associated with each tweet, we get to explore the effect of the structures of this popular social network on the diffusion of opinions. A naive initial approach would be compiling a list of popular topics from news articles and searching for the occurrence of these keywords in each tweet. We show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. Using the corpus, we build a sentiment classifier that can determine positive, negative and neutral sentiments for a document. We propose to use comments gathered from Yelp! as labeled training data for this project. We expect that the meaning of phrase to context relation is roughly preserved across texts generated through online user input and that the Yelp! comments would be able to capture most of the trending phrases.

WHY R?

R is the leading tool for performing data analysis operations such as virtual data manipulation, statistical model, and charts. It is more than a statistical package; it’s a programming language, so you can create your own objects, functions, and packages. You can easily use it anywhere.  It’s [platform-independent](http://en.wikipedia.org/wiki/Platform-independent), so you can use it on any operating system. R allows you to integrate with other languages (C/C++, [Java](http://java.sun.com/) [and Python](http://www.python.org/)) and enables you to interact with many data sources: [ODBC](http://en.wikipedia.org/wiki/Odbc)-compliant databases (Excel, Access) and other statistical packages ([SAS](http://www.sas.com/)).

FEATURES OF R

R is more than just a domain-specific programming language aimed at data analysis. It has some unique features that make it very powerful. They are

1. R PACKAGES

R is a plethora of libraries that include generic R Libraries focusing on general purpose such as econometrics, natural language processing, etc., and particular R libraries that specialize on industry problems (such as clinical trials, genetics, finance.

1. PERFORMING MULTIPLE CALCULATIONS WITH VECTORS

R is a vector-based language. You can think of a vector as a row or column of numbers or text. The list of numbers {1, 2, 3, 4, 5}, for example, could be a vector. Unlike most other programming languages, R allows you to apply functions to the whole vector in a single operation without the need for an explicit loop.

1. PROCESSING MORE THAN JUST STATISTICS

As R started to expand away from its origins in statistics, many people who would describe themselves as programmers rather than statisticians have become involved with R. The result is that R is now eminently suitable for a wide variety of non-statistical tasks, including data processing, graphical visualization, and analysis of all sorts. R is being used in the fields of finance, natural language processing, genetics, biology, and market research, to name just a few. R is Turing complete, which means that you can use R alone to program anything you want.

1. RUNNING CODE WITHOUT A COMPILER

R is an interpreted language, which means that — contrary to compiled languages like C and Java — you don’t need a compiler to first create a program from your code before you can use it. R interprets the code you provide directly and converts it into lower-level calls to pre-compiled code/functions.

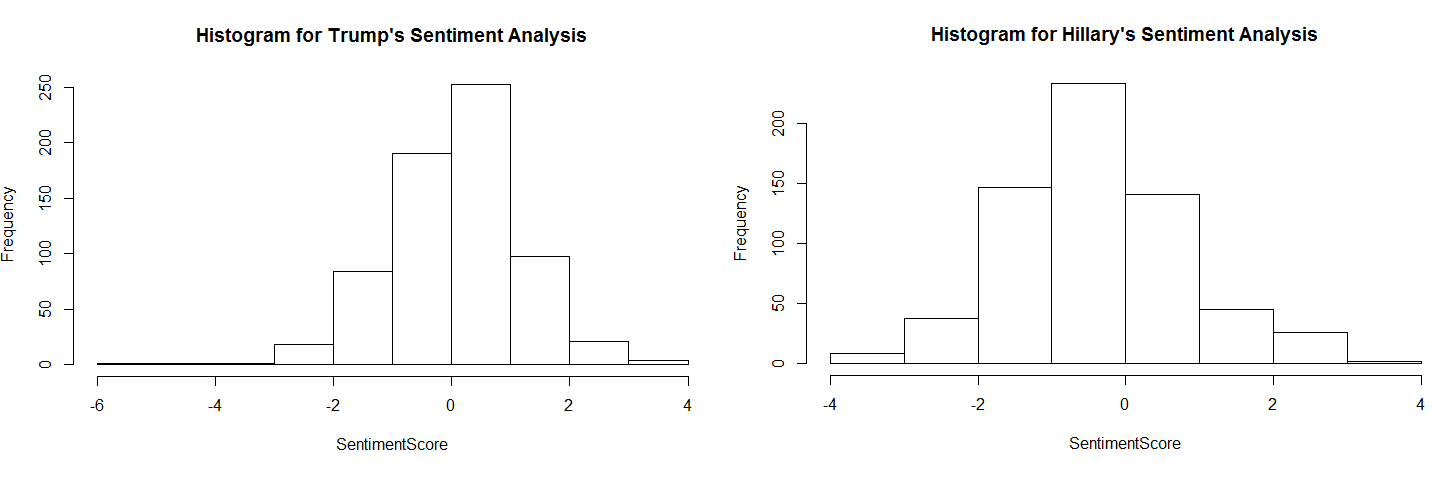
1. R IS FREE AND OPEN SOURCE SOFTWARE

It allows anyone to use and, importantly, to modify it. R has no license restrictions (other than ensuring our freedom to use it at our own discretion), and so we can run it anywhere and at any time, and even sell it under the conditions of the license.

USAGE OF R IN THE PROJECT:

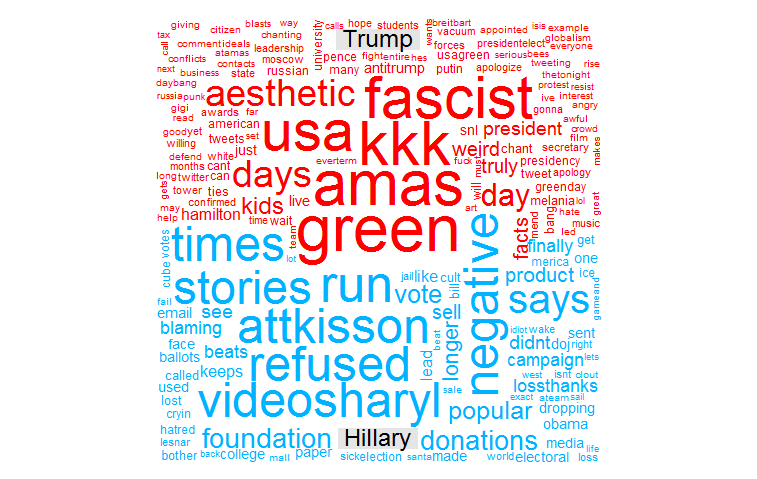
* **TwitteR** package in R can be used to do a lot of analysis on social media “Twitter”. We have used the searchTwitter function to search twitter on a search-term and load the tweets that are saved to a variable for further processing.
* In order to analyze the tweets, we first extracted the text from tweets and saved it into the variable. We converted the list of tweets to a vector using sapply function of **Plyr** package. Plyr allows the user to split a data set apart into smaller subsets, apply methods to the subsets, and combine the results.
* We then cleaned the text by removing unnecessary characters, emoticons and numbers by making a call to custom built function clean.text( )
* Load the positive and the negative wordlist file. Scan the file for Positive words and Negative words and store it locally.
* We used sentiment.R for evaluating the sentiment. The function sentiment.score( ) takes three arguments the text to be analyzed, positive words and negative words.
* The above function converts the sentences in the text to words using the split methods and retrieve the appropriate scores using score methods. The score is calculated by counting the number of occurrences of positive and negative words in a tweet.
* The analysis results can be used to draw table using built-in table( ) function or histogram can be plotted by using R’s built-in hist( ) function.

Below screen shots capture the sentiment score results for tweets on Donald Trump and Hillary Clinton.



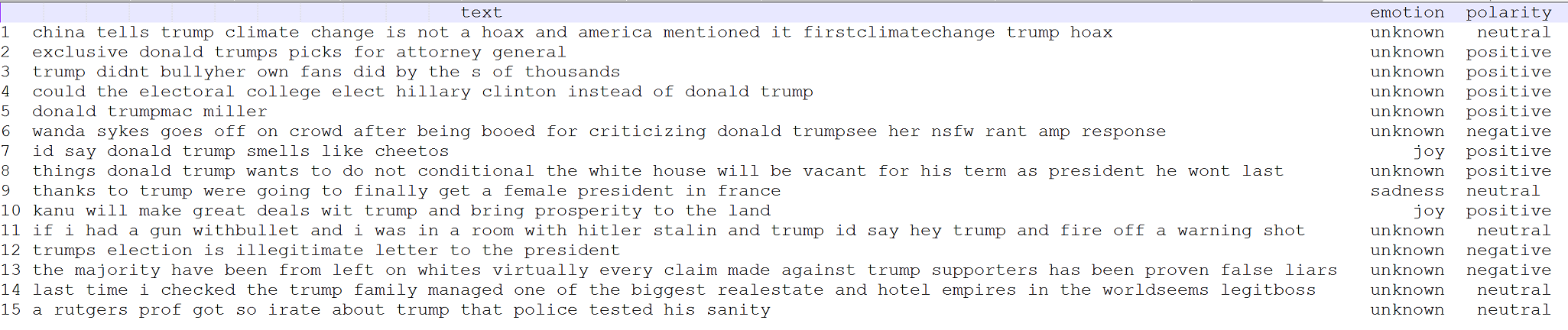
The above analysis shows that most of the response about Donald Trump is positive when compared to that of Hillary Clinton.

A comparison cloud drawn by taking the top 500 words from tweets of both Trump and Hillary:

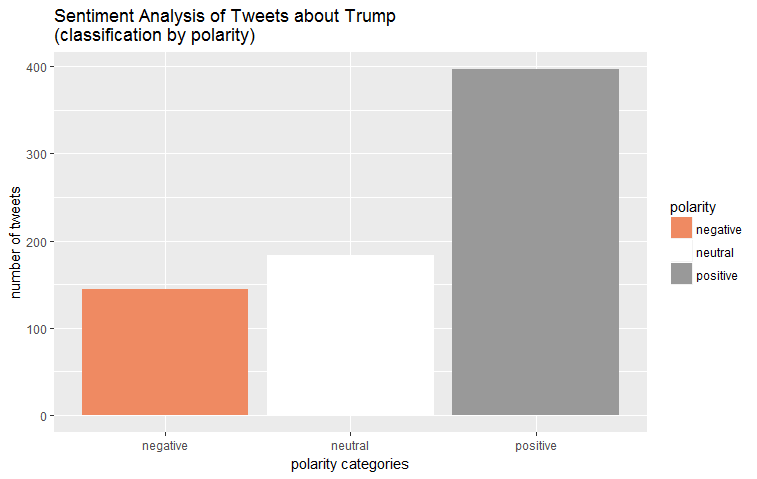


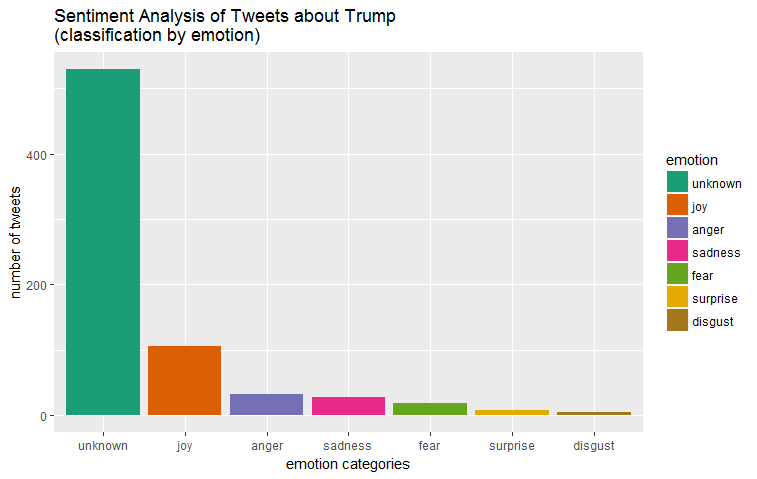
* By using the **sentiment** package of R, we can also analyze the text and classify it based on emotion (anger, disgust, fear, joy, sadness, and surprise) or polarity. Performing sentiment analysis on cleaned text of Donald Trump’s tweets yielded the following results.

Top 15 tweets after running the algorithm of sentiment analysis:



**Plot distribution of the obtained results classified by polarity and emotion are depicted below**





**Classification of the text by emotions and visualization with a comparison cloud**

