Twitter Data Classification   
By Haritha Chinta, Gulsum Oz

Neeta Gupta, Samit Ashdhir

Table of Contents

**The Problem2**

**Objective 2**

**Method**

**Analysis 3**

Word Cloud 3

Data Partition 4

KNN Model5

Decision tree Model6

**Model comparison** 7

**Conclusion** 8

The Problem

Evaluate methods to classify twitter streams using machine learning techniques

The Objective

This projects aims to classify twitter data by using text mining techniques and machine learning algorithms. Many machine learning algorithms are implemented to see performance.

The Methods

For this project first we collected 20000 mentions from twitter half belongs to Hillary Clinton and the other half belongs to Donald Trump. Next we used text mining method to clean the data and transformed words and sentences into a form that a computer can understand. Later we gave a score to each tweet based on opinion word usage in each tweet. Than we labelled each tweet positive or neutral. After visualizing the data we implemented classification with different machine learning algorithm. We used "class", "MASS", "ggplot2", "tm", "twitteR", "dplyr", packages in R to accomplish the project.

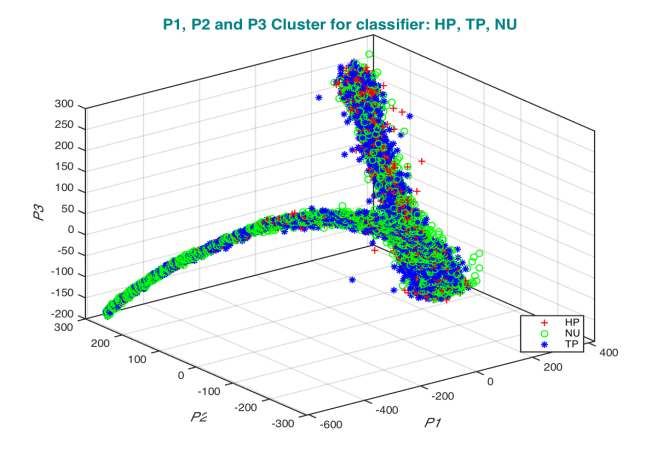
The Analysis

First step toward constructing the classifier involves processing the raw data for analysis. We collected twitter data by using twitter developer account. Data is basically a text document labelled as either Hillary Clinton or Donald Trump. Twitter mentions are string of text composed of words, spaces, numbers, and punctuations. We formed a corpus from the text document which refers to a collection of text document. In this project a text document refers to a single twitter mention.

Then we used (tm) package in R to handle the text cleaning. We removed punctuations, stop words such as “and” , “but”, “or” and extra white spaces between words. Now our data is ready for analysis.

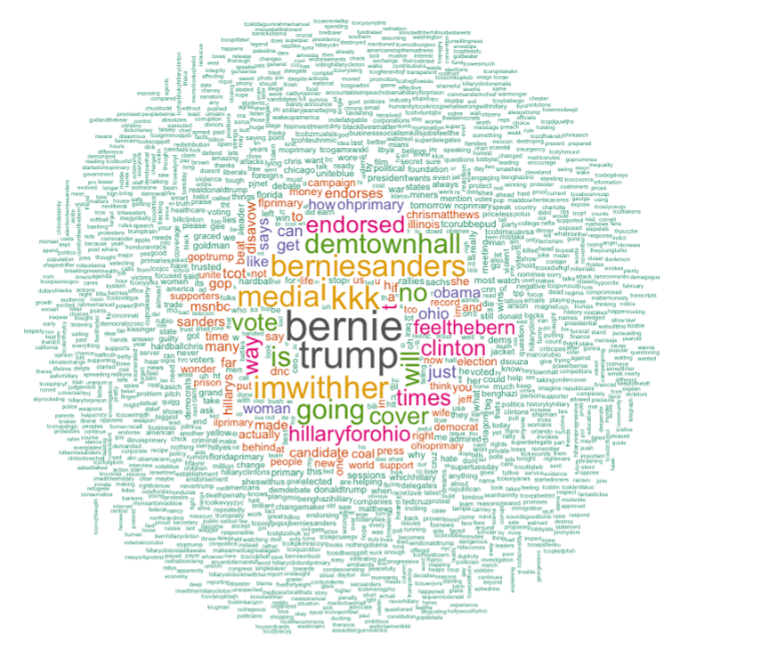
Most of tweets contains opinion words such as good, great, exciting, bad, awful etc… comparing each mentions with opinion lexicon in English, computer gave a score for each tweet. Now each tweet is labelled positive or neutral based on its score.

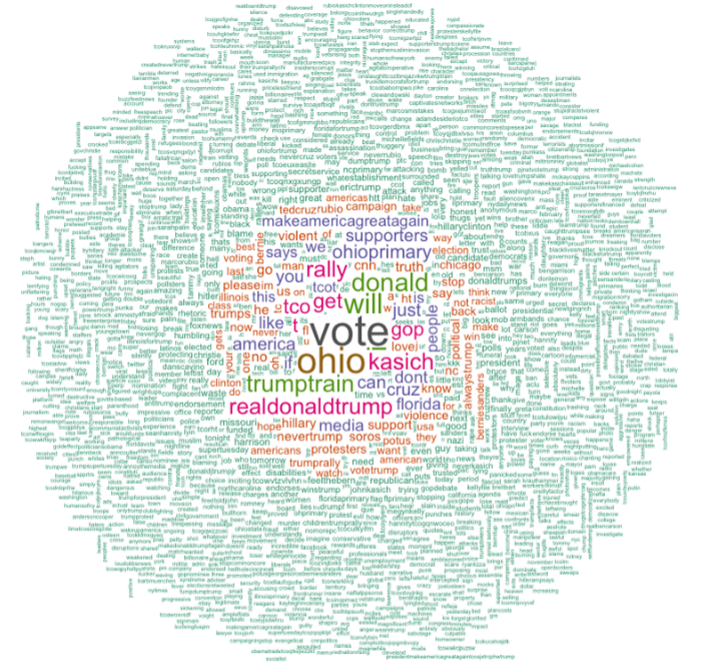
Next step we converted each tweet strings to number so that computer can handle it. Below is the plot of the first three PC’s after the dimension reduction.



Word Cloud

Our data now ready for the classification but, before doing anything we tried to visualize the data by forming word clouds. We formed separate word cloud for Hillary Clinton and Donald Trump to see the frequent word usage for both candidates.





Data Partition

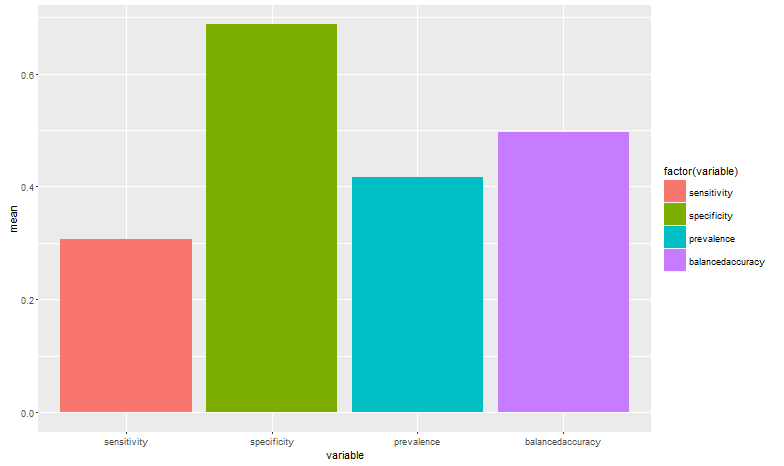
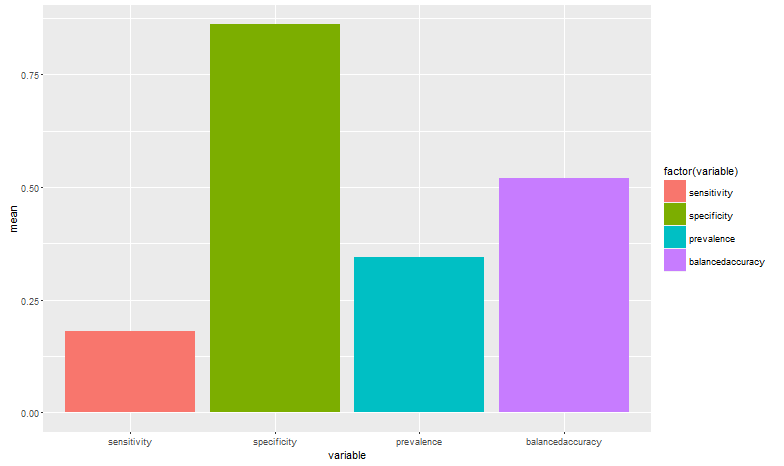
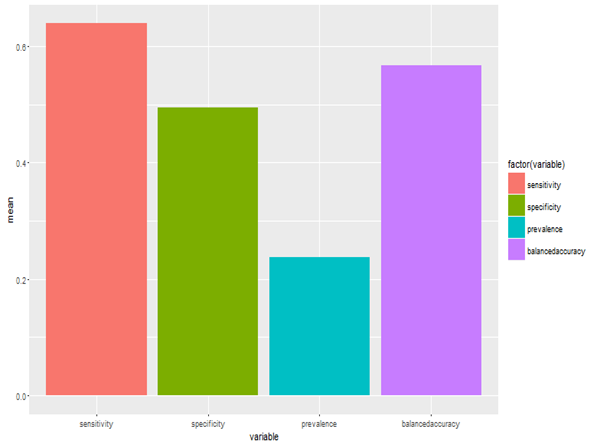
As a next step we randomize the data and divided data into training and the test set by %75 training and %25 test partition ratio.

Histogram Classifier

Linear Classifier

Naïve Bayes Classifier

In this type of classifier R built-in function used to perform the classification. The entire dimension is used to do this classification. Naïve Bayes is known for the text classification but in our project it didn’t perform well. Accuracy was 0.3415. Below are plots performances for each class

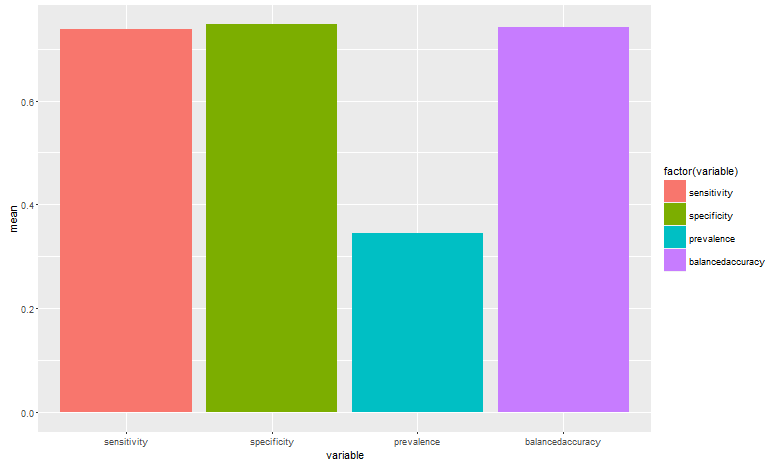
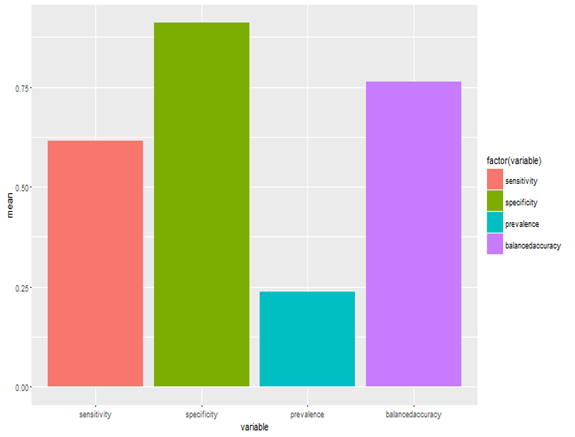


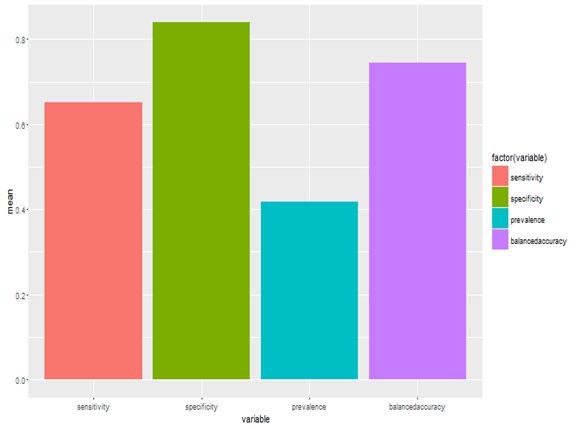
(Performances: left top Class HP, Left right Class NU, Left below Class TP)

K-NN Classifier

In this classifier (class) package in R used to perform the classification. K-NN known as the slower algorithm for performing the classification due to the distance calculation between each data point, I think we had a chance to see this. Data dimension was so huge for the K-NN to perform the classification so we took 1/3 of the training data and accomplished K-NN classification. Accuracy was close to the decision tree classification. Below are the performance plots for KNN model.

Accuracy: 0.6721





(Performances: Left top Class HP, Right Top Class NU, Left below Class TP)

Conclusion

Over all, this project taught us a lot. First of all we saw how growing text mining area applications. Second it was nice to see different machine learning algorithm implementation and their performances. We had a chance to work on a big dimensional data like text body of the tweets and learned how to handle it with application in R. Results of the classifications relatively satisfying for us and we will continue digging more to explore new ways to get better resuts.