STA 141A Final Project

Siddhanth Sabharwal, 999229332 John Nguyen, 998808398

06/06/2017

Honor Code: The codes and results derived by using these codes constitute my own work. I have consulted the following resources regarding this assignment:

None

Introduction

The data set chosen is a Twitter gender classification data set. The data set has information for 20,050 users across 26 variables. For these users the actual gender is given along with a probability level that the information is correct. Before training classification algorithms on this data, some data cleanup is required. First, all non-human users such as brands are removed. Second, users that can't be determined to be humans are removed. Third, any user where Twitter wasn't 100% confident about the gender was excluded so to not incorrectly train our algorithm.

Goals and Questions

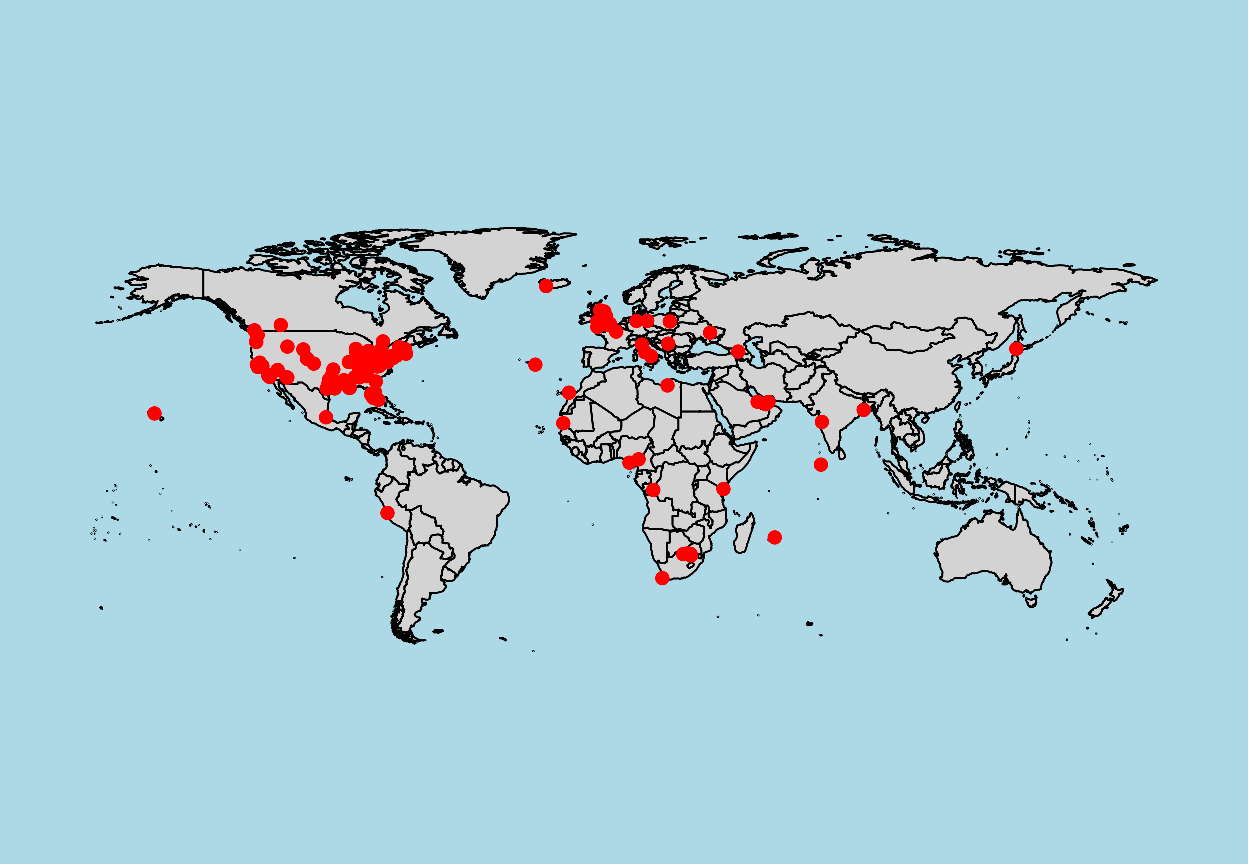
For our project, we are trying to answer the following questions:

1. How well do words in tweets and profiles predict user gender?
2. What are the words that strongly predict male or female gender and what are the most popular words among them?
3. How well do stylistic factors (like link color and sidebar color) predict user gender?
4. How can we classify gender based on the user descriptions?

Data

Each row has a user name, a random tweet, account profile and image, location, and even link and sidebar color. The data Twitter User Gender Classification used for the project is available on Kaggle website. The data table includes 26 columns and 20,051 rows. Each row in the data represent a Twitter users, and the columns are are the data collected on the user. We also gather a data set of 100 common male and female names in the US to perform regular expression search on the username in order to deepen our classfication model for gender. The data is available on Social Security website.

Where our data points are coming from.



The column labels are shown below:

```
##  [1] "X_unit_id"      "gender"         "created"        "description"
##  [5] "fav_number"     "link_color"     "name"           "retweet_count"
##  [9] "sidebar_color"  "text"           "tweet_coord"    "tweet_count"
## [13] "tweet_location" "user_timezone"
```

Methods

For classification of the gender data, we use kNN and Naive Bayes methods. We also use graphs and plots to visualize the data. Using text column, we also find the frequency of top common words used in tweets based on gender and perform a clustering by Hierarchical Clustering using complete linkage

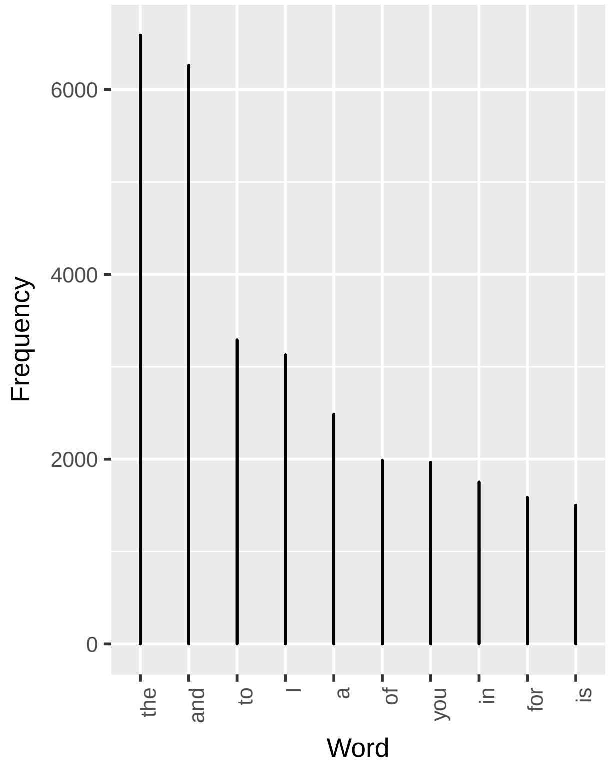
Gender Classification

Naive Bayes Classifer

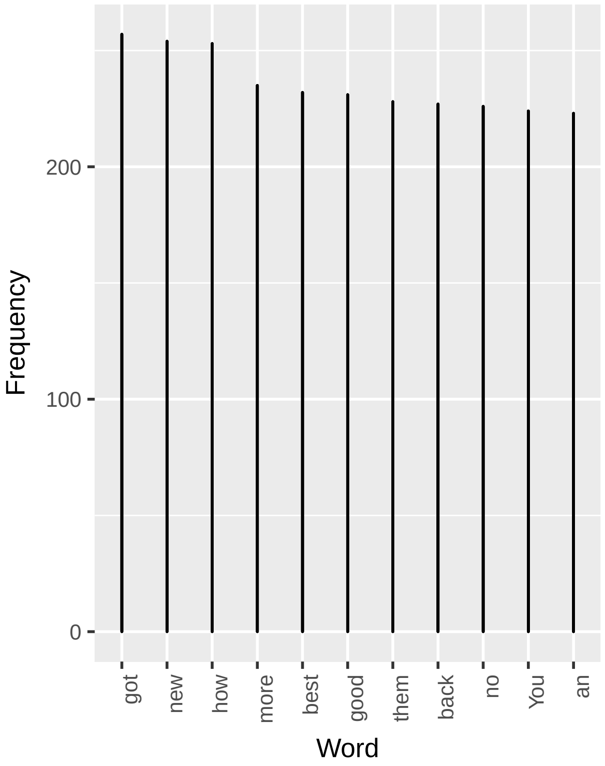
Most Commonly Used Words Among Genders

Unsurprisingly the top 10 words are all propositional phrases. Therefore I took a step further to filter out these words in order to get a deeper learning about the sentiment on twitter.

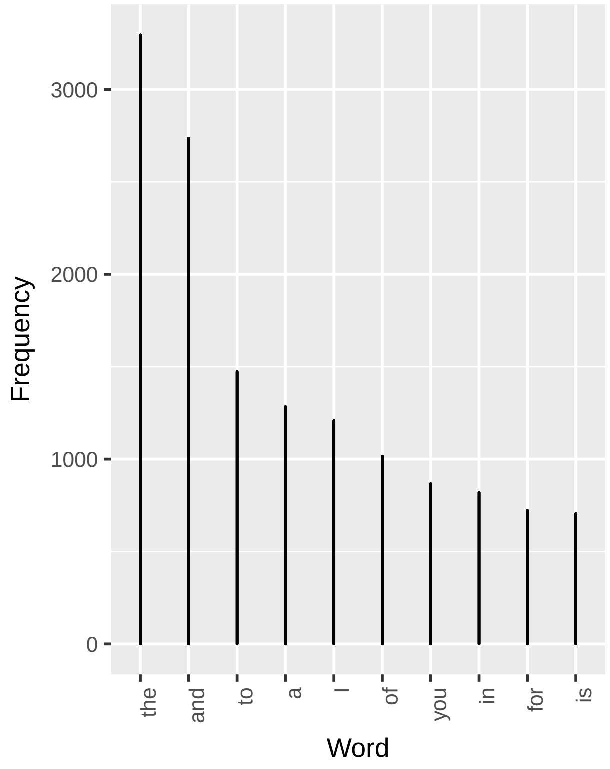
Top 10 Words
by both gender.



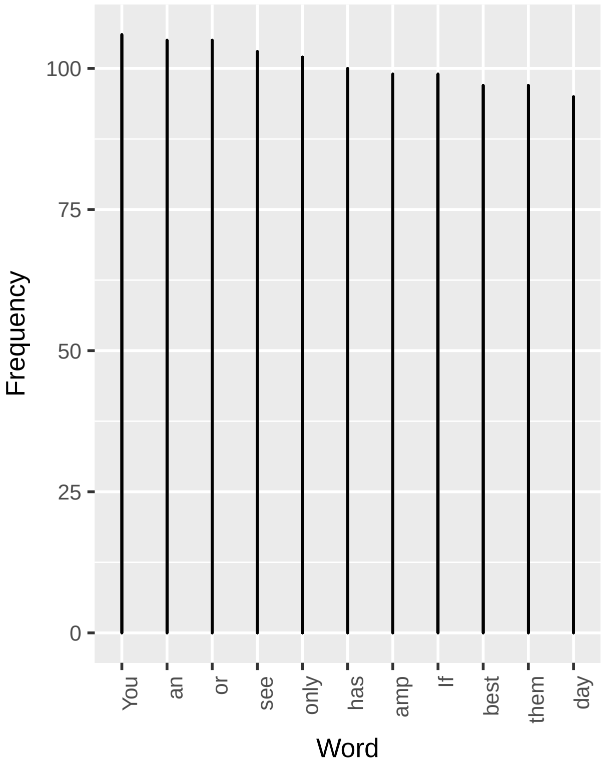
Top 10 Non−propositional Words
by both gender.

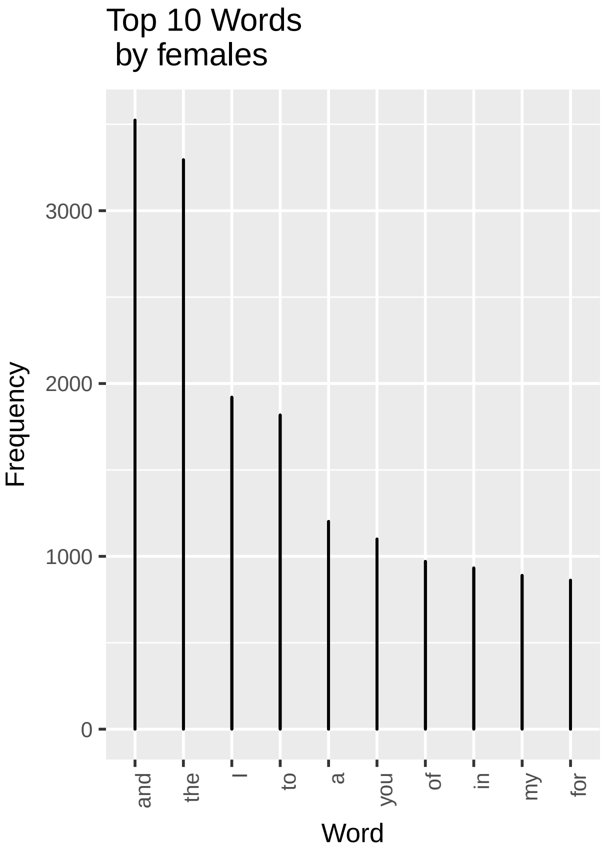


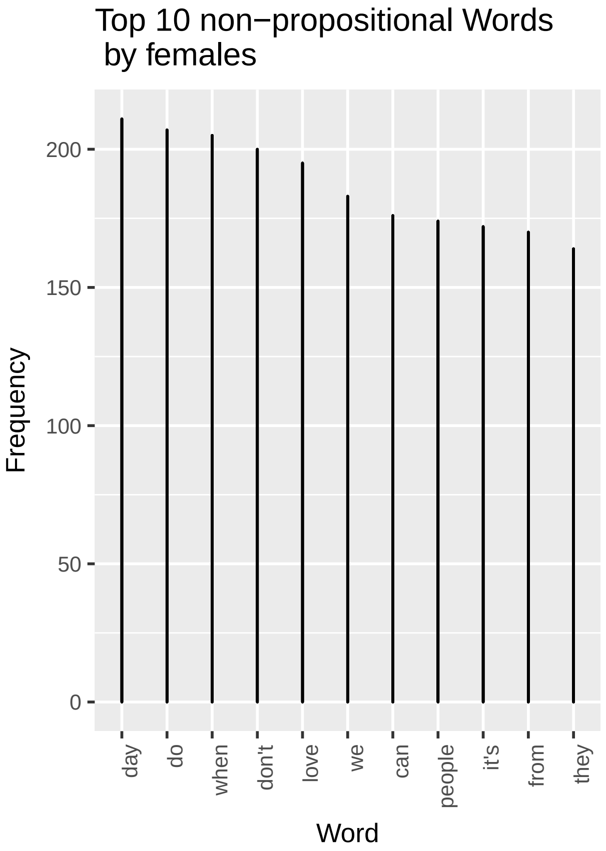
Top 10 Words
by males



Top 10 Non−propositional Words
by males





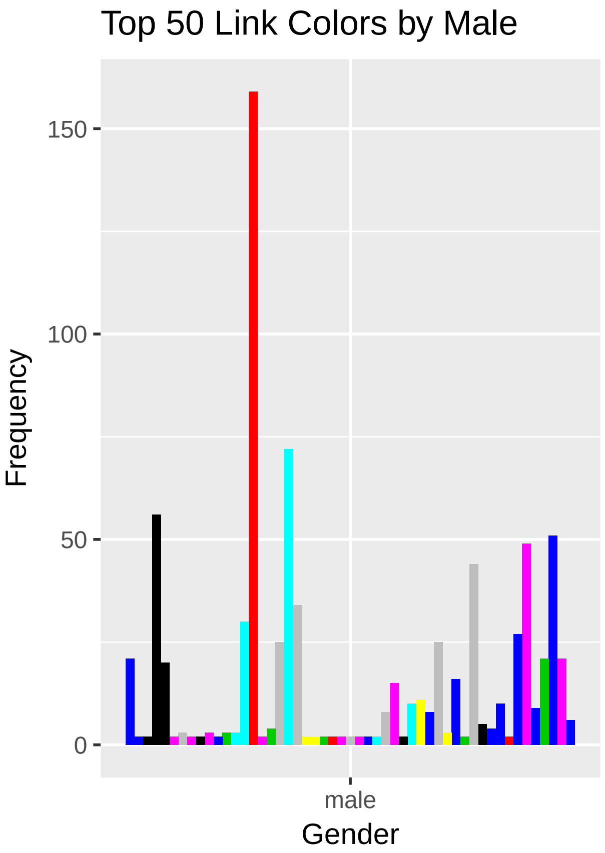


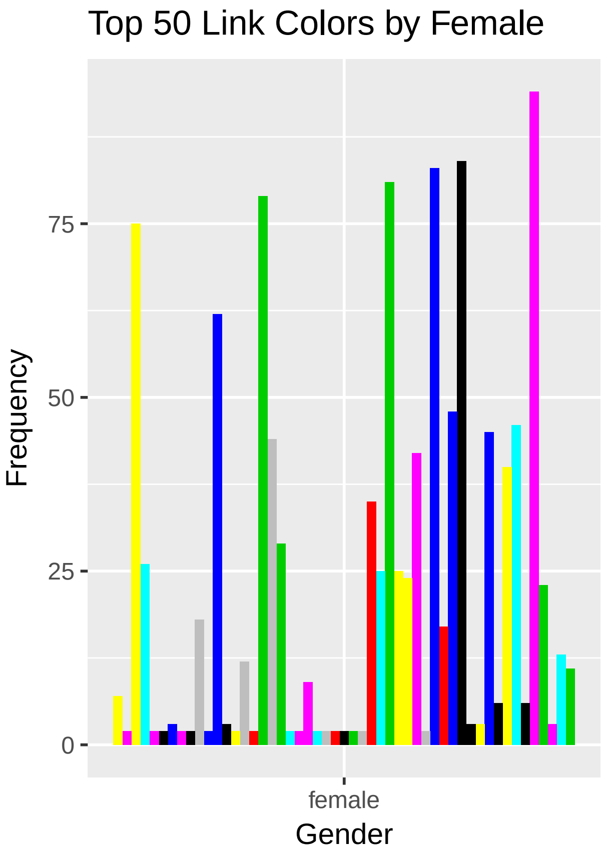
These are the top 10 words that are said my males but not by females.

```
## [1] "team"     "games"    "top"      "support"  "Bond"     "See"
## [7] "brothers" "tax"      "Well"     "Will"
```

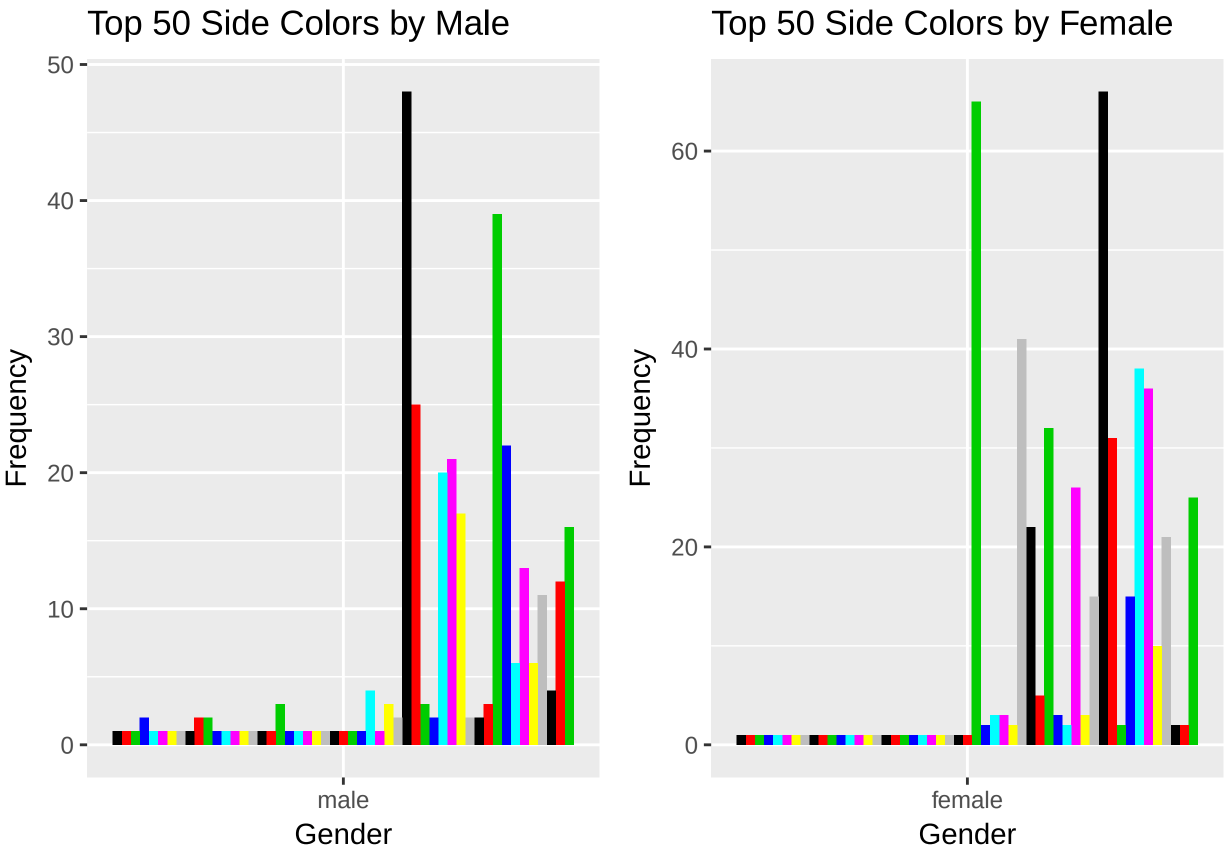
As one can see, these more masculine words, for example *brothers* or *Bond* are more commonly used by males. *Bond* here is referring to the James Bond movies,generally more popular among males. We can come to a generally conclusion that these words can help us identify a person's gender and that these are the words that strongly predict male or female gender.

Stylistic factors In Predicting User Gender





The subset of data we are interested for in this case refers to the link_color and side_bar color columns. For this reason, we extract this information and create data frames into tables using the ggplot function. These values are tabulated into a bar graph with a pictoral display of color by gender. We'll ignore the values for brand, and notice that 0084B4 is the most common color for females,as given by its hex value. For males, this hex value is 08C2C2. These hex values 0084B4, and 08C2C2 represent orange, and red, respectively. The hex vaue 0084B4 has the highest frequency. Females used this color nearly 2500 times for their link color, and it lead the other color choices by a considerable margin. The second occurrence of the most used color is the red, 08C2C2, by males which was used close to 500 times. Since there was such a large discrepancy and difference amongst usage of orange as by females, and by red for males as link profile colors, it can be assumed that link color of a profile can be particularly useful when it comes to predicting gender.



We perform a similar anlaysis with sidebar color as we did with link color. Again we use ggplot to generate a pictoral display of the various colors used for the sidebar, broken down by gender. At first glance, this distribution looks quite different, however it is still clear that there are preferred colors for each gender. The largest frequency in colors occurs from brands that tweeted, so we can ignore this. In general the highest frequency of color used amongst genders was considerably less than that of link color. The hex value 087FA7, corresponding to the color red, was the most commonly used color for the sidebar for both males and females. This color was used over 400 times amongst females and 500 amongst males. However, because this particular color was the most favorited amongst both genders, it would not give any meaningful insight when it comes to predicting gender. Although there are several predictors and various conditional probabilites, this would be a weak indicator of gender when considered solely on its own.

Conclusion

Through an analysis of of the metrics of classification and identification for tweets, it was important to first predict, and then confirm which variables were meaningful in predicting gender. In our case, we limited our analysis to words, stylistic factors such as link color and profile sidebar color, as well as user descriptions.

We ignored common propositional phrases which would, by default, be most common amongst the tweets and came to find that there were certain words and phrases which we deemed as more "masculine". Conversely, there were words that could be heavily attributed to females.The idea is that certain words and phrases may be linked to the collective gender's experiences. Common masculine words such "Bond" and "brother" reinforce this point. Because there was no overlap amongst common words and phrases said by males and females, if we are excluding propositional phrases, we conclude that we have some basis for predicting gender based on word choice.

Another metric of evaluating tweets involved the overall stylistic and visual appearance. We came to find that link color and sidebar color were unrelated to each other as their distributions were entirely different based on gender. While orange was far and away the most common link color for females, males preferred

red, but at a considerably lesser extent. While choice of these two colors for links could be indicative of gender, it is unclear whether any inferencs could be made based on gender for the other colors. The sidebar color distribution did not provide much information that could lead to a prediction for gender. Both males and females preferred red when it came to this sidebar color.

Packages Used

For our analysis, we use the R programming language including libraries: ggplot2,dplyr,stringr, RTextTools, naivebayes, sets, data.table

References

1.