MODULE 4- TWITTER SENTIMENTAL ANALYSIS

Kiran Khanderao - X160128 *Date: 3/1/2019*

OBJECTIVE

This assignment is an illustration of classification based on **Naïve Baye** prediction model using **Twitter Text** data to identify whether the sentiment expressed in the messages is “Positive” or “Negative”.

NAÏVE BAYE – PREDICTION MODEL / TWITTER SENTIMENTAL ANALYSIS

Step 1: The dataset has 1.6 million records which needs to cut down in size after randomizing using

Step 2: “Exploring and preparing the data” to pre-process data set to a “bag of words”

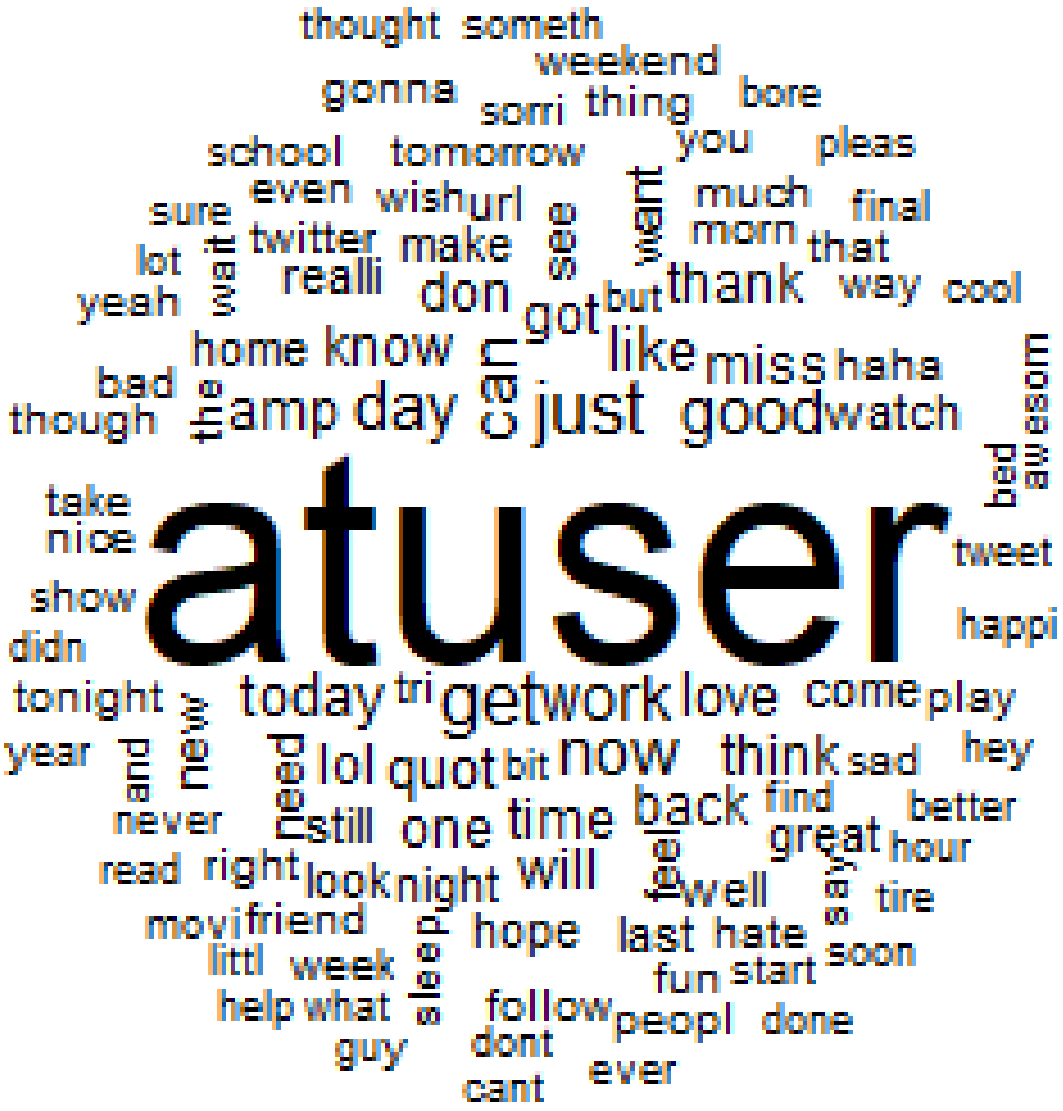
Step 3: The “Target” element is a character vector (“Positive” or “Negative”). Since this is a categorical variable, it would be better to convert it into a factor followed by **Data preparation – cleaning and standardizing text data.**

Step 4: The first step in processing text data involves creating **a corpus**, which is a collection of text documents using Vcorpus() function in tm package. The we clean up data and remove filler words (stop words) followed by “Remove Punctuation” and finally stemming (reducing to the root word).

Step 5: Data Preparation: Now we reduce cleaned data by splitting text documents into wordspredict (p) based with the final step to split the messages into individual components through a process called **tokenization** resulting in a data structure called a Document Term Matrix (DTM) in which rows indicate documents (text messages) and columns indicate terms (words).

Step 6: Next, we create training and test datasets and test it along with Visualizing text data – word clouds and training a model on the data, evaluating model performance.

Step 6: The “Cross Table” function shows



THE RESULT

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  1000
```

```
##
##
##               | actual
##    predicted  |  negative |  positive | Row Total |
## -------------|-----------|-----------|-----------|
##     negative |       328 |       132 |       460 |
##              |     0.671 |     0.258 |           |
## -------------|-----------|-----------|-----------|
##     positive |       161 |       379 |       540 |
##              |     0.329 |     0.742 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       489 |       511 |      1000 |
##              |     0.489 |     0.511 |           |
## -------------|-----------|-----------|-----------|
##
##
```

INTERPRETATION OF THE RESULTS

Note 1: This is not a large sample data set but just 1000 records.

Note 2: the two classifications of the text data are “Positive” and “Negative” only.

Note 3: Out of the 460 actual negative data, 67.1% was identified as -ve. (True Negative) and 25.8% was identified as positive “False Positive”

Note 4: Out of the 540 actual positive data, 74.2% (Approx.) was identified as +ve. (True Positive) and 33% was identified as positive “False Negative”

Note 5: Totally there were 51% positive comments and 49% negative comments