
# Tweet Analysis & Sentiments Prediction

30.04.2018

**Submitted by :**


| Name | Enrollment Number | Batch |
|---|---|---|
| Parimal Mishra | 15103147 | B2 |
| Vishal Malhotra | 15103167 | B2 |
| Kushagra Bindra | 15103301 | B7 |


**Submitted To :** Dr. Adwitya Sinha

## Abstract :

In today's world where social media usage is growing exponentially, there exist a need to organize, analyze and predict the contents shared over it. Twitter which receives almost 6000 tweets per seconds and has a sum total of 232 million active users has become one of the fastest growing social platform for information sharing and awareness. The usage of Twitter isn't only restricted to news and entertainment, but it is being extensively used for communicating problems, reporting issues, open discussions, creating awareness etc. Therefore while keeping all these features of twitter in mind, the project, Tweet Analysis and Sentiment Prediction in R language tries to frame the data of twitter in more structured way by applying Data Preprocessing over it. The processed data is then visualized in all possible respects using the Data Visualization techniques available in R. Along with all the visualizations, there also exist plotting of the location on the world map from where the tweet has been made by a user given another dimension to this project and data visualization. This makes the collected data more effective and impactful. And thereafter implementation  of several Machine Learning algorithms like KMeans, SVM, Naive Bayes and Decision Tree to find the relevant pattern hidden within the data and predicting the possible outcomes on certain instances. There also has been work done on the future aspects and possibilities of the project.

## Introduction :

Tweet Analysis and Sentiment Prediction is a project which mainly focuses on the techniques of Data Preprocessing, Data Visualization and Machine Learning algorithms with the future scope of implementation of real time extraction of informative measures from a dataset and Database Connectivity for the future references. This project had several different dimension of each categories mentioned above ranging from data retrieval through twitter API to filtration of data i.e. removal of unnecessary details in Data Preprocessing, plotting count of variables to geographical location plotting in Data Visualization along with the use of Shiny package for plotting sentiment histograms, pattern finding to predicting the possible responses and outcomes in Machine Learning. The challenges associated with the project included the learning and implementation of R language in all the three categories, followed by the objective of working upon the pre existing datasets available in the public domain and then the compilation of entire project under one package. Upon implementation of the of all the techniques mentioned above the project had a well developed model which is capable enough of making the raw data made available ready for effective use and visualization. After this the project model can be utilized to dig up some existing data patterns hidden inside the dataset and find the relation amongst different variables, collect some most informative information and can later on can be trained well enough to predict the response and conclusion on a particular or a group of inputs using the machine learning techniques. The future aspects upon implementation will make the model more effective and stable on the grounds of data analysis and sentiment prediction.

# Model Assumptions :

- **Data Source and Dataset :**
  - **Data Source** : https://www.kaggle.com
  - **Dataset** : Sentiment classification of tweets
    - ➢ Number Of Rows : 14641
    - ➢ Number Of Columns : 13
  - **DatasetLink**:https://www.kaggle.com/valencar/sentiment-classification-of-tweets

- **Tools/Techniques/Libraries :**
  - **Tools :**
    - ➢ **R Language**
    - ➢ **RStudio**
  - **Techniques :**
    - ➢ **Data Preprocessing**
    - ➢ **Data Visualization**
    - ➢ **Machine Learning**
    - ➢ **Geographical plotting**
  - **Libraries :**
    - ➢ **Rtexttools :** RTextTools is a machine learning package for automatic text classification.
    - ➢ **E1071 :** Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier
    - ➢ **Dplyr :** dplyr is an R package for working with structured data both in and outside of R. dplyr makes data manipulation for R users.
    - ➢ **Tidytext :** In this package, we provide functions and supporting data sets to allow conversion of text to and from tidy formats, and to switch seamlessly between tidy tools and existing text mining packages.

- **Nbclust :** NbClust package provides 30 indices for determining the number of clusters and proposes to user the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods.
- **Tm:**A framework for text mining applications within R.
- **snowballC :** An R interface to the C libstemmer library that implements Porter's word stemming algorithm for collapsing words to a common root to aid comparison of vocabulary.
- **Wordcloud :** Use for generating pretty word clouds.
- **Rcolorbrewer:** Provides color schemes for maps (and other graphics) designed by Cynthia Brewer.
- **Stringr:** Character manipulation,Whitespace tools to add, remove, and manipulate whitespace,Locale sensitive operations,Pattern matching functions.
- **Corrgram:** Calculates correlation of variables and displays the results graphically. Included panel functions can display points, shading, ellipses, and correlation values with confidence intervals.
- **Corrplot:**The package is a graphical display of a correlation matrix, confidence interval.
- **Plotrix:** This R package provides lots of plots, various labeling, axis and color scaling functions.
- **httr:** The aim of httr is to provide a wrapper for the curl package, customised to the demands of modern web APIs. ... Response content is available with content() as a raw vector ( as = "raw" ), a character vector ( as = "text" ), or parsed into an R object ( as = "parsed" ), currently for html, xml, json, png and jpeg .
- **Rmongo:** MongoDB Client for **R**. MongoDB Database interface for **R**. The interface is provided via Java calls to the mongo-java-driver.
- **Rworldmap:** Enables mapping of country level and gridded user datasets.
- **Ggplot2:** ggplot2 allows you to create graphs that represent both univariate and multivariate numerical and categorical data in a straightforward manner.
- **twitteR:** provides access to the Twitter API. Most functionality of the API is supported,with a bias towards API calls that are more useful in data analysis as opposed to daily interaction.
- **Shiny:** makes it easy to build interactive web apps straight from R. You can host standalone apps on a

webpage or embed them in R Markdown documents or build dashboards.

- **Methods and Algorithms :**
  - **KMeans**
  - **SVM**
  - **Naive Bayes**
  - **Decision Tree**
  - **Word Cloud**

# Experimental Outcomes :

- **Experimental Values :**

  - **KMeans :**
    - Number of clusters : 30
    - Size of the clusters:

      68, 99, 150, 3561, 161, 56, 5, 236, 877, 231, 202, 99, 96, 242, 402, 76, 123, 258, 269, 499, 196, 276, 1051, 1019, 92, 177, 305, 367, 4, 3443

    - Cluster means :

| | ASC | NRC |
|---|---|---|
| 1 | -1.6852010 | -2.278550e+00 |
| 2 | -1.1883076 | -2.278550e+00 |
| 3 | -1.6596998 | -1.785996e-02 |
| 4 | 0.6127283 | -1.090208e-03 |
| 5 | -1.2103922 | 2.303399e-01 |
| 6 | -1.1708156 | -9.591346e-01 |
| 7 | 0.5725675 | 5.625950e-01 |

| | | |
|---|---|---|
| 8 | -3.3488103 | -1.938172e+00 |
| 9 | 0.6115159 | -1.006216e+00 |
| 10 | -1.3480853 | 1.512204e-01 |
| 11 | -1.5230372 | -1.061206e+00 |
| 12 | -1.1933448 | 1.442934e-04 |
| 13 | -1.6382575 | -1.097729e+00 |
| 14 | -1.3132790 | -2.278550e+00 |
| 15 | -1.5390621 | 1.439233e-02 |
| 16 | -1.3904991 | -9.592591e-01 |
| 17 | -1.3381843 | -1.029793e+00 |
| 18 | -1.4551625 | 8.898087e-02 |
| 19 | -1.3310513 | 3.963192e-16 |
| 20 | 0.6131006 | 2.217691e-01 |
| 21 | -1.4333970 | -1.047949e+00 |
| 22 | -1.4848735 | -2.278550e+00 |
| 23 | 0.6129008 | 7.569384e-02 |
| 24 | 0.6131006 | 1.422074e-01 |
| 25 | -1.2665610 | -9.752702e-01 |
| 26 | -1.5762583 | -2.278550e+00 |
| 27 | -1.4004673 | -2.278550e+00 |
| 28 | -1.4353478 | 3.963192e-16 |
| 29 | 0.2578220 | -2.278550e+00 |
| 30 | 0.6128211 | 1.290746e+00 |

- **SVM :**
  - Number of Support Vectors : 14
  - Coefficients:

    (Intercept) y = -0.2318 x = 0.9487
  - Parameters:

    SVM-Type: eps-regression

    SVM-Kernel: radial

    cost: 1

gamma: 1

epsilon: 0.1

- Parameter tuning of ‘svm’:
- sampling method: 10-fold cross validation

- **Naive Bayes :**
  - Training Set Size : 10981
  - Test Set Size : 3660
  - Accuracy : 80 %

- **Decision Tree :**
  - Training Set Size : 22
  - Test Set Size : 7
  - Classes : positive,negative,neutral
  - Predicted

- **Word Cloud :**
  - Dataset size : 14641
  - Most used words :
    - flight
    - unit
    - usairways
    - americanair
    - southwestair

## ❖ RESULTS :

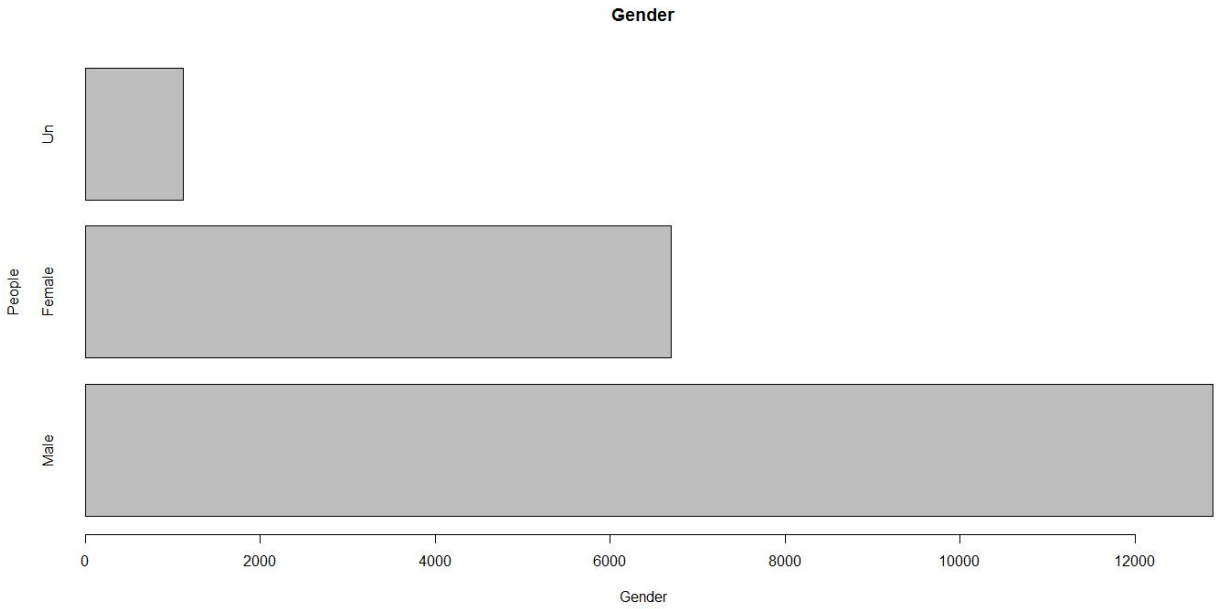



**Fig 1. Plot between Gender and People.**

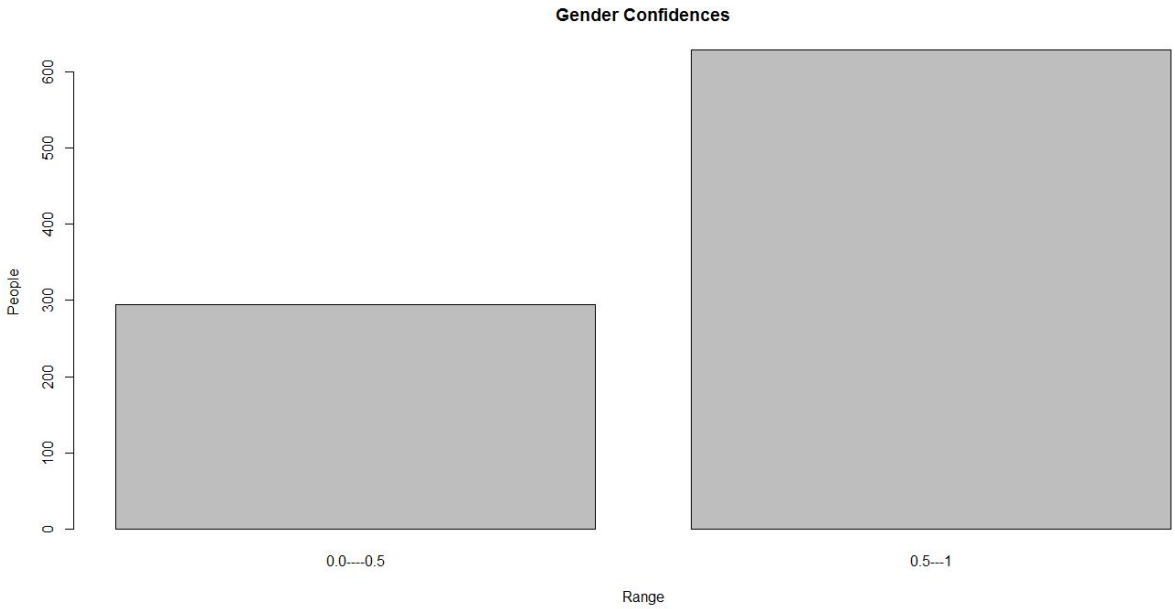



**Fig 2. Plot of Gender Confidence.**

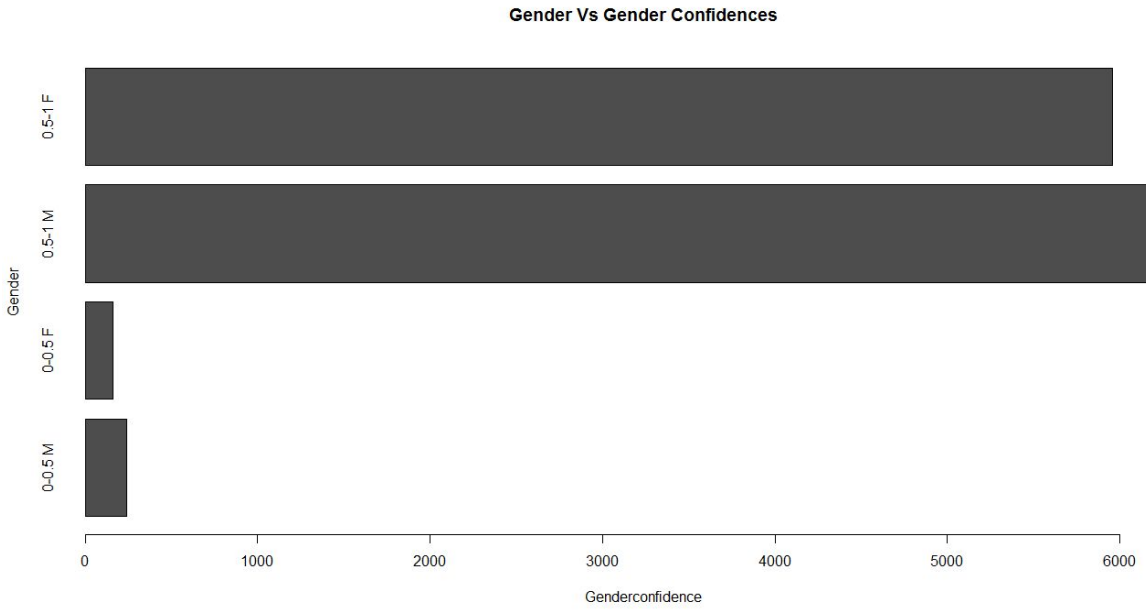



**Fig 3. Plot for Gender vs Gender Confidence.**

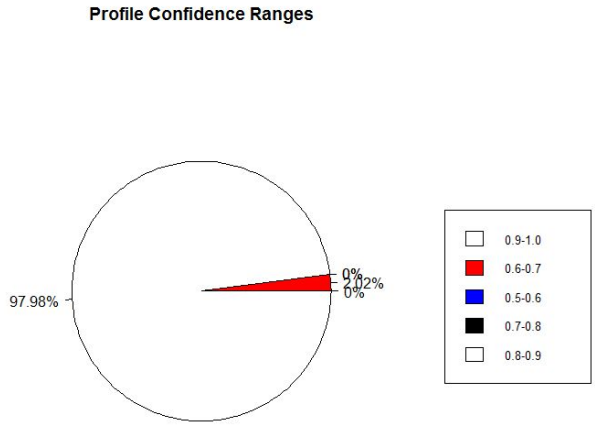



**Fig 4. Pie chart for Profile Confidence**

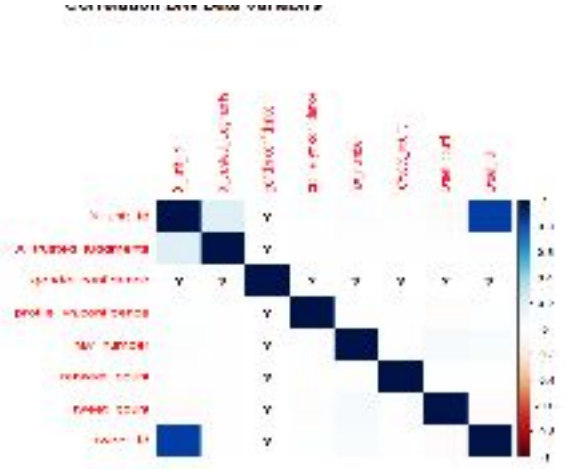


**Fig 5 . Correlation Measure**

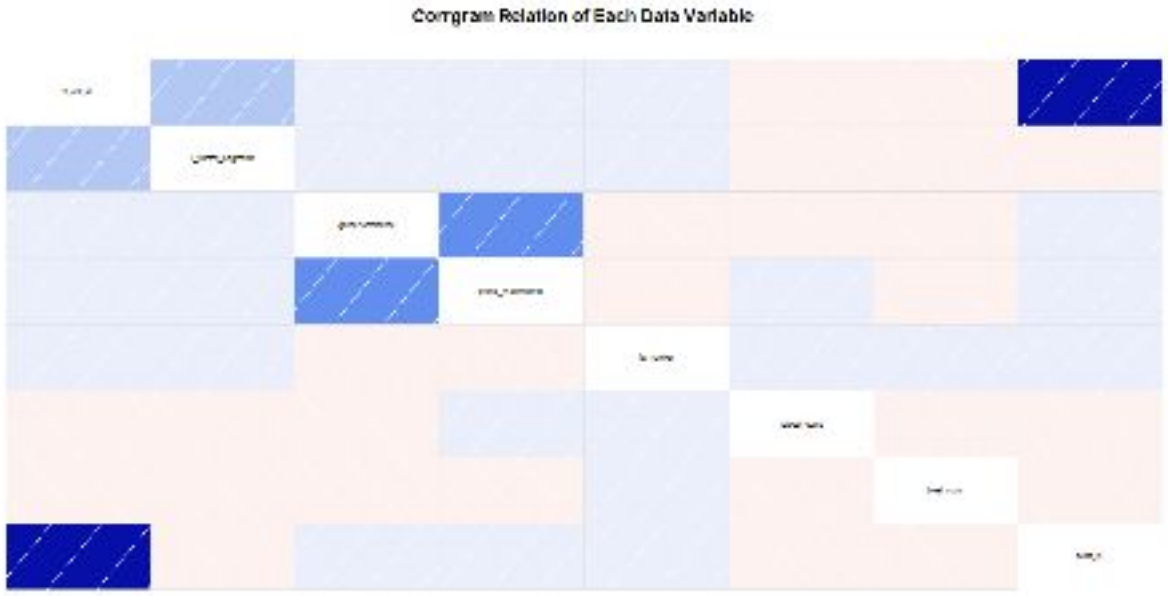



**Fig 6. Corrgram Relation 1**

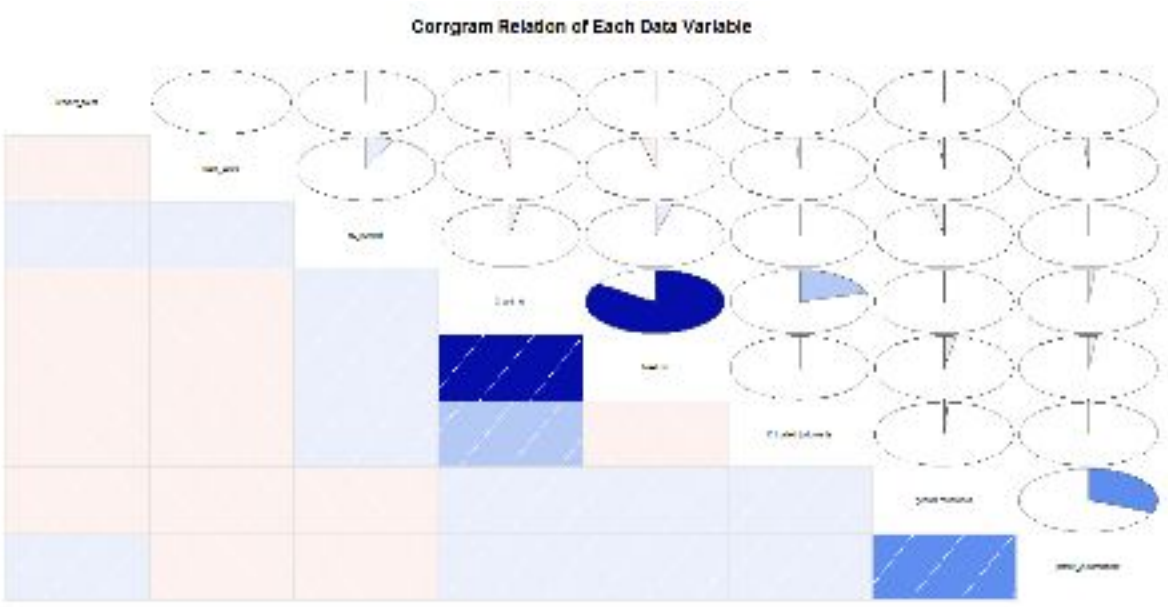



**Fig 7. Corrgram Relation 2**

- **K-Means:**

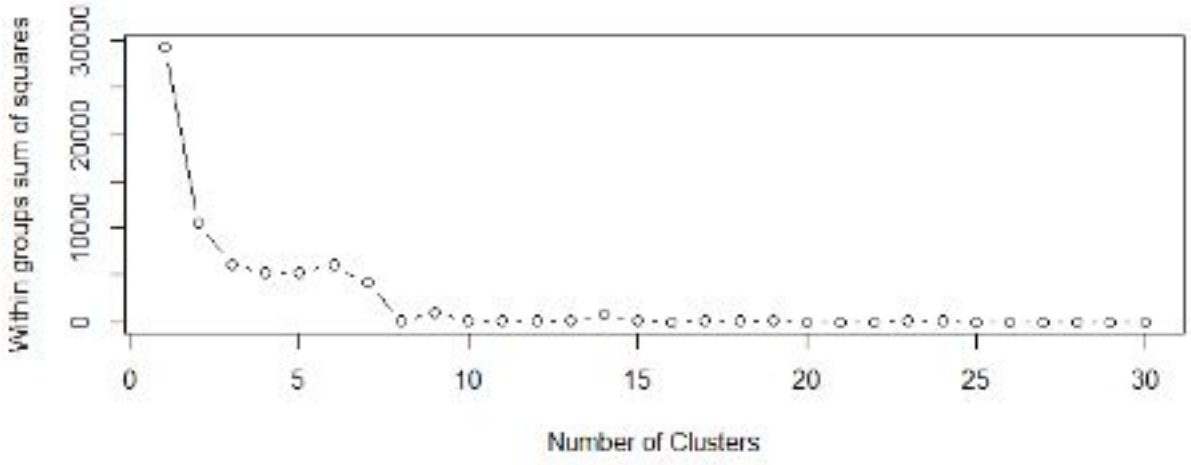



**Fig 8 . K-Means Plot**

- **SVM :**
  - ❖ best parameters:

    epsilon = 0.07  cost = 256
  - ❖ best performance: 0.03784239

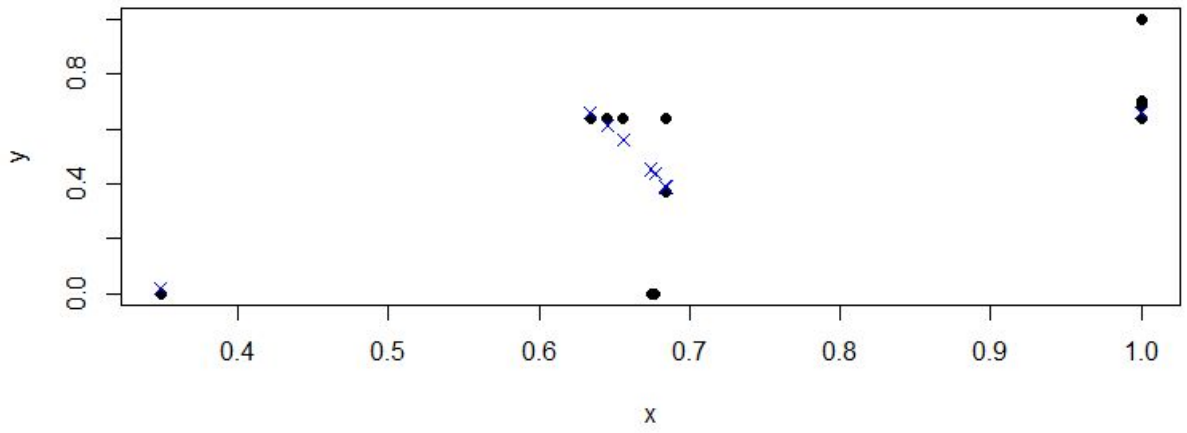



**Fig 9 . SVM Performance Measure**

- **Naive Bayes:**

**Rows : Actual outcome**

**Column : Predicted outcome**

| | negative | neutral | positive |
|---|---|---|---|
| negative | 3 | 0 | 0 |
| neutral | 0 | 5 | 0 |
| positive | 0 | 2 | 0 |

- **Decision Tree :**

  **Rows : Testing tweet number**

  **Columns : Possibility of tweets**

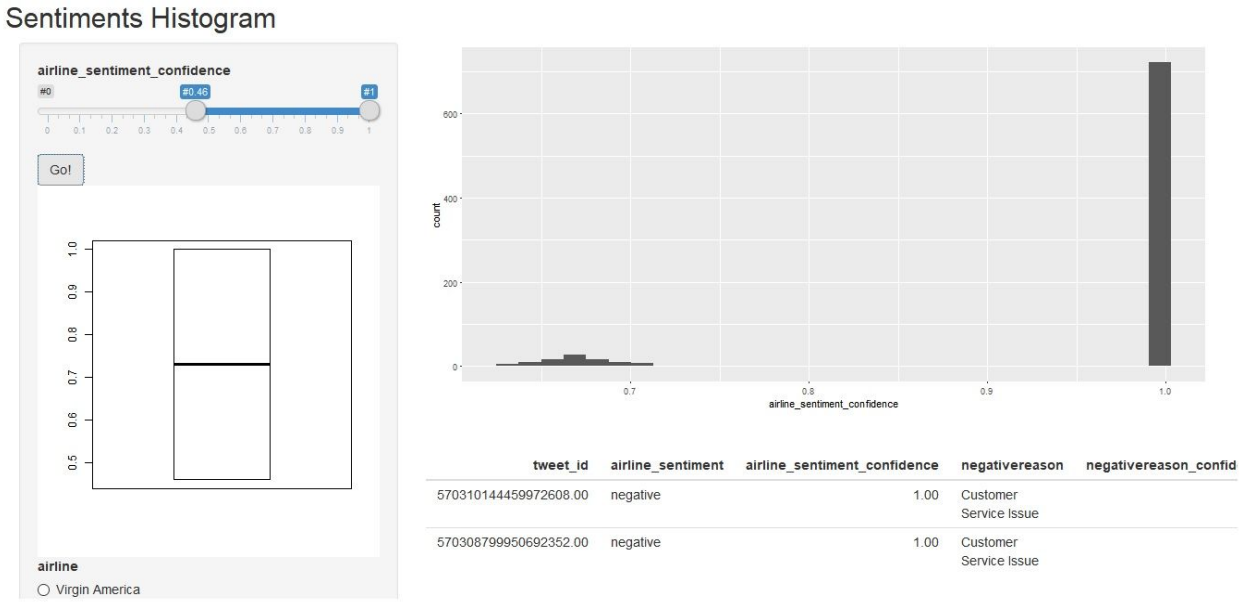
| | Negative | Neutral | Positive |
|---|---|---|---|
| 1 | 0.5000000 | 0.0000000 | 0.5000000 |
| 2 | 0.5000000 | 0.0000000 | 0.5000000 |
| 3 | 0.5000000 | 0.0000000 | 0.5000000 |
| 4 | 0.5000000 | 0.0000000 | 0.5000000 |
| 5 | 0.1428571 | 0.7142857 | 0.1428571 |
| 6 | 0.5000000 | 0.0000000 | 0.5000000 |
| 7 | 0.5000000 | 0.0000000 | 0.5000000 |
| 8 | 0.5000000 | 0.0000000 | 0.5000000 |

- **Word Cloud :**

  - Dataset size : 14641
  - Most used words :
    - flight
    - unit
    - usairways
    - americanair
    - southwestair

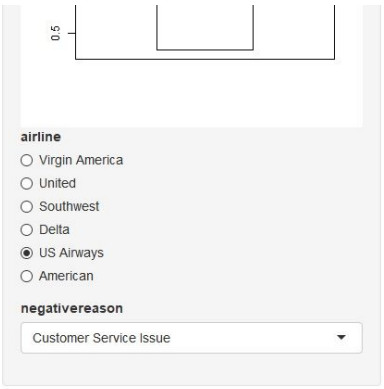

**Fig 10(a) . Word Cloud**


**Fig 10 (b). Word Cloud**




| tweet_id | airline_sentiment | airline_sentiment_confidence | negativereason | negativereason_confid |
|---|---|---|---|---|
| 570310144459972608.00 | negative | 1.00 | Customer Service Issue | |
| 570308799950692352.00 | negative | 1.00 | Customer Service Issue | |

**Fig 11(a) . Histogram for sentiment analysis.**




| tweet_id | airline_sentiment | airline_sentiment_confidence | negativereason | negativereason_confid |
|---|---|---|---|---|
| 570310144459972608.00 | negative | 1.00 | Customer Service Issue | |
| 570308799950692352.00 | negative | 1.00 | Customer Service Issue | |
| 570307605631012864.00 | negative | 1.00 | Customer Service Issue | |
| 570304779412508672.00 | negative | 1.00 | Customer Service Issue | |
| 570304445336178688.00 | negative | 1.00 | Customer Service Issue | |
| 570304440017788928.00 | negative | 1.00 | Customer Service Issue | |
| 570303720308809728.00 | negative | 1.00 | Customer Service Issue | |
| 570302717085790208.00 | negative | 1.00 | Customer Service Issue | |
| 570300325917270016.00 | negative | 1.00 | Customer | |

**Fig 11(b) . Histogram for sentiment analysis.**

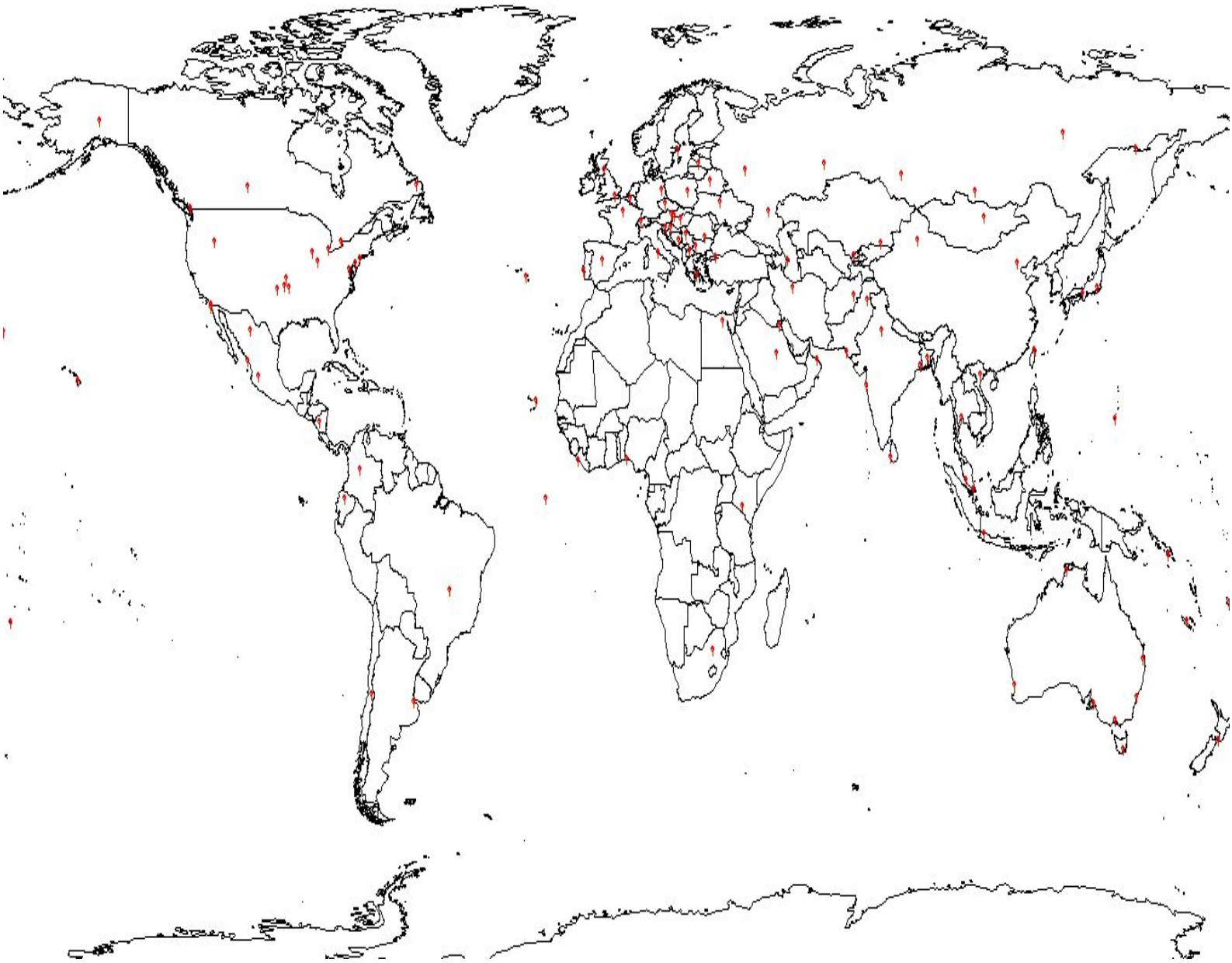


**Fig 12 . World Map of the Locations Frequent for Tweets.**

# Conclusion :

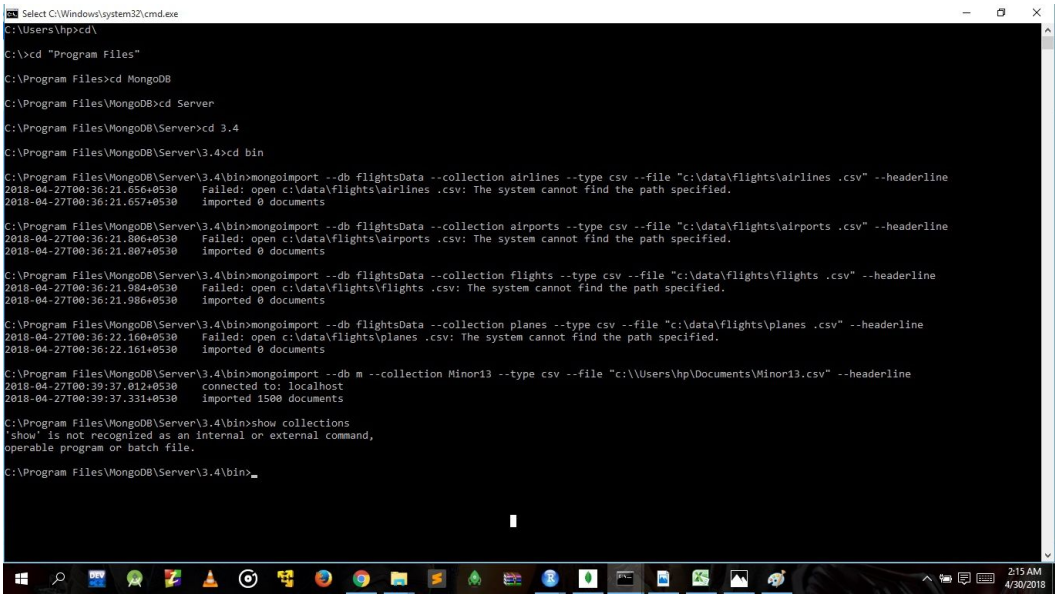
- Conclusive Remarks :
  Tweet Analysis and Sentiment Prediction is a project which mainly focuses on the techniques of Data Preprocessing, Data Visualization and Machine Learning algorithms

  ➔ Field and techniques worked upon :

  - Data Preprocessing:
    - Dataset collection
    - Data refining (Noise reduction, NA values etc)
    - Data filtration
  - Data Visualization:
    - Plots
    - Histograms
    - Geographical Plotting
  - Machine Learning:
    - KMeans
    - SVM
    - Naive Bayes
    - Decision Tree
    - Word Cloud
  - Work Contribution :

| S. No. | Name | Enrollment | Batch | Field of Working |
|---|---|---|---|---|
| 1. | Parimal Mishra | 15103147 | B2 | Machine Learning, Data Preprocessing |
| 2. | Vishal Malhotra | 15103167 | B2 | Data Visualization, Data Preprocessing, Future Aspect |
| 3. | Kushagra Bindra | 15103301 | B7 | Data Visualization, Data Preprocessing, Future Aspect |

- **Future Work Directions :**
  This project upon worked ahead can lead to the following achievements :
  - ❖ Making the prediction real time.
    - ➢ Use of the twitter API to fetch the tweets in real time.
    - ➢ Feeding the data into the several models leading to analysis and prediction .
  - ❖ Use of database to maintain the records.
    - ➢ Database handling platforms like MongoDB can be used in order to store the data along with the results for the future reference.
    - ➢ Records available of the previous findings might be useful in witnessing the pattern followed by the phenomenon.By phenomenon here, we are referring to several different situations like Tsunami, Earthquake, Flood, Diseases etc.



# References:

[1] M. Sarnovsky , P. Butka , A. Huzvarova , “Twitter data analysis and visualizations using the R language on top of the Hadoop platform” , *IEEE 15th International Symposium on Applied Machine Intelligence and Informatics,* vol. 15, pp. 327-331, 2017.

[2] M. Bashri , R. Kusumaningrum, “Sentiment Analysis Using Latent Dirichlet Allocation and Topic Polarity Wordcloud Visualization”, *2017 Fifth International Conference on Information and Communication Technology (ICoICT),* vol. 5, 2017.

[3] N. Garg, R. Rani, “Analysis and Visualization of Twitter Data using k-means Clustering”, *International Conference on Intelligent Computing and Control Systems ICICCS 2017,* 2017.

[4] P. Manivannan, Dr. P. Isakki, Dengue Fever Prediction Using K-Means Clustering Algorithm”, *2017 IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT TECHNIQUES IN CONTROL, OPTIMIZATION AND SIGNAL PROCESSING,* 2017.

[5] Introduction to visualising spatial data in R Robin Lovelace (R.Lovelace@leeds.ac.uk), James Cheshire, Rachel Oldroyd and others 2017-03-23.

[6] Wickham, H. (2014). Tidy data. The Journal of Statistical Software, 14(5), Retrieved from http://www. jstatsoft.org/v59/i10 .

[7] Cheshire, J., & Lovelace, R. (2015). Spatial data visualisation with R. In C. Brunsdon & A. Singleton (Eds.), Geocomputation (pp. 1–14). SAGE Publications.