SENTIMENT ANALYSIS ON TWEETS FOR AIRLINES

Harshit Srivastava (19BCE0382)

School of Computer Science and Engineering

Vellore Institute of Technology

Abstract: This project, is based on textual information processing has been focused on mining and retrieval of factual information, e.g., information retrieval, Web search, text classification, text clustering and many other Data Mining concepts. In this project we work on a dataset that contains tweets related to various airlines based in the United States of America.

I. Introduction

The problem of sentiment analysis: As for any scientific problem, before solving it we need to define or to formalize the problem. The formulation will introduce the basic definitions, core concepts and issues, sub-problems and target objectives. It also serves as a common framework to unify different research directions. From an application point of view, it tells practitioners what the main tasks are, their inputs and outputs, and how the resulting outputs may be used in practice.

In a large-scale network, for example, an online informal organization, we could have a large number of hubs and edges. Identifying people groups in such systems turns into a troublesome assignment. In this way, we need network identification calculations that can segment the system into numerous networks.

II. Methodology

Data Cleaning: Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Data Visualisation: Data visualization is an interdisciplinary field that deals with the graphic representation of data. It is a particularly efficient way of communicating when the data is numerous as for example a Time Series. From an academic point of view, this representation can be considered as a mapping between the original data (usually numerical) and graphic elements.

Data Mining: Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information from a data set and transform the information into a comprehensible structure.

III. Analysis

Word cloud: A word cloud is an assortment, or group, of words portrayed in various sizes. It is another type of plotting the information, helps understanding the plot with the words itself. The bigger and bolder the word shows up, the more frequently it's referenced inside a given book and the more significant it is. Word mists can be plotted in numerous shapes, for example, circle, oval, square, and so forth.

Social Network diagram: A social network diagram outwardly shows the connections and collaborations between hubs (individual entertainers, individuals, or things inside the system). It maps out the hubs (people or gatherings) and the ties, edges, or links (relationships or communications) that interface them. The connection between two hubs characterizes the degree to which hub structure attaches with comparable versus different others.

It can be arranged into bunches based on classifying a gathering of people who communicate with each other and offer comparative interests. Identifying such types of clusters is known as community detection.

Community detection: Community detection is the identification of groups that are densely connected with nodes. The community detection algorithms used in this experiment are:

- Estimate Betweenness detection algorithm.
- Label propagation algorithm.
- Greedy optimization algorithm.

Clustering: K-Means and Hierarchical Clustering algorithms will be used to analyse Negative and Positive Tweets in our Dataset.

- **k-means clustering** is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances.

- **Hierarchical clustering**, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. Hierarchical clustering can be performed with either a distance matrix or raw data. When raw data is provided, the software will automatically compute a distance matrix in the background.

IV. Literature Review

- With the movement of web advancement and its turn of events, there is a tremendous volume of data present in the web for web customers and a lot of data is created too.

- Individual to individual correspondence goals like Twitter, Facebook, and so on are rapidly getting as they license people to share and express their points of view about subjects, have discussion with different systems, or post messages over the world.

- Twitter starting at now gets around 190 million tweets every day, where people share their comments regarding a wide extent of focuses.

- A gigantic number of tweets fuse opinions about things and organizations.

- The episode of PC based feeling investigation just happened with the accessibility of abstract messages on the Web.

- Numerous subjects past item surveys like securities exchanges, decisions, fiascos, medication, programming designing and cyberbullying broaden the use of slant examination.

- Label propagation is a community detection technique where each node possesses a label denoting the community to which it belongs. A node acquires the label given to most of its neighbouring nodes.

- Prior knowledge about communities is not necessarily needed in label propagation-based algorithms.

- A greedy algorithm is an algorithm that follows the problem-solving heuristic of making the locally optimal choice at each stage with the intent of finding a global optimum.

- In many problems, a greedy strategy does not usually produce an optimal solution, but nonetheless a greedy heuristic may yield locally optimal solutions that approximate a globally optimal solution in a reasonable amount of time.

- Estimate Edge Betweenness as a Centrality Measure.

V. Data Visualisation

For data visualisation in R, we use libraries like ggplot2, plotly, wordcloud, tm, dplyr, maps, rbokeh, htmlwidgets, widgetframe, etc.

Different Graphs for various categories have been plotted.

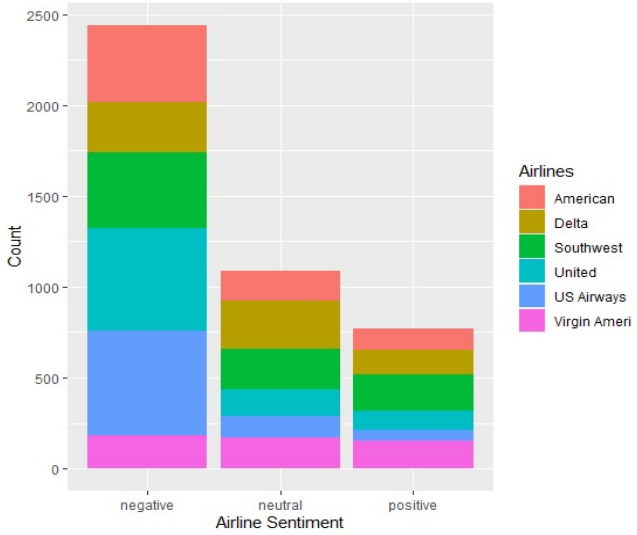

Figure 1: Airline Sentiment

In Figure 1, We see the counts of tweets with respect to negative, positive and neutral sentiments according to different airlines.

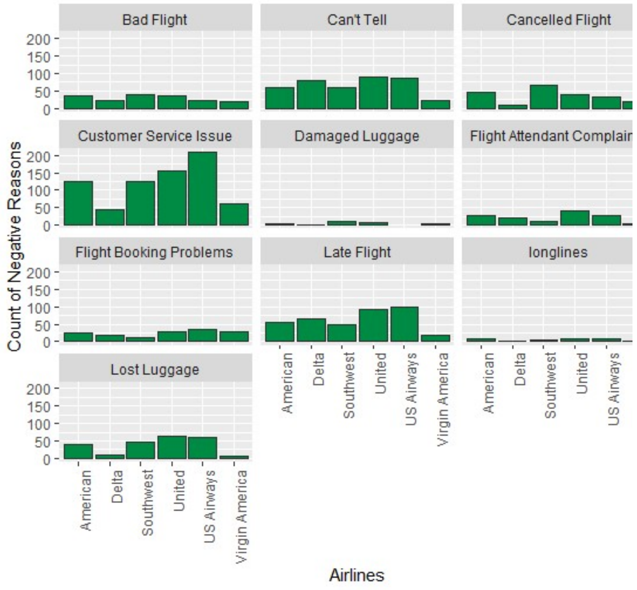

Figure 2 Count of Reasons

In Figure 2, We can see the count of tweets frequency for negative tweets and the reasons for such tweets with respect to each Airline.

Word Clouds:

+ Negative and Positive word clouds depict frequency of words used in that context.

+ In negative cloud words like 'cancelled' are prominent whereas in positive cloud words like 'thanks' are prominent.

+ Word clouds with respect to each Airline show the context of words most used for those airlines.

+ A combined word cloud shows the most frequent words in tweets throughout the dataset.

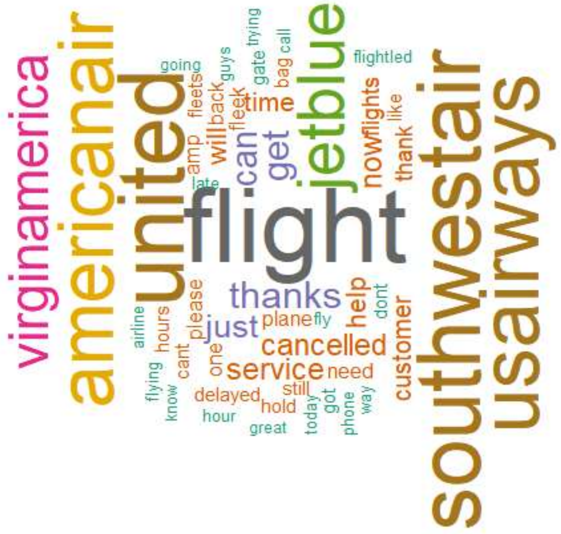

Figure 3 Word Cloud

In Figure 3, We can see that words like 'flight', airline names, 'cancelled', 'thanks', etc. This signifies the frequency of these words used in tweets throughout the dataset. The words larger in size are the words more frequently used than the ones with smaller size.

VI. Sentiment Analysis

Syuzhet Package: The package comes with four sentiment dictionaries and provides a method for accessing the robust, but computationally expensive, sentiment extraction tool developed in the NLP group at Stanford. Use of this later method requires that you have already installed the coreNLP package.

get_sentiment(): After you have collected the sentences or word tokens from a text into a vector, you will send them to the get_sentiment function which will asses the sentiment of each word or sentence. This function takes two arguments: a character vector (of sentences or words) and a "method." The method you select determines which of the four available sentiment extraction methods to employ. In the example that follows below, the "syuzhet" (default) method is called.

The get_nrc_sentiment function returns a data frame in which each row represents a sentence from the original file. The columns include one for each emotion type was well as the positive or negative sentiment valence. The example below calls the function using the simple twelve sentence example passage stored in the s_v object from above.

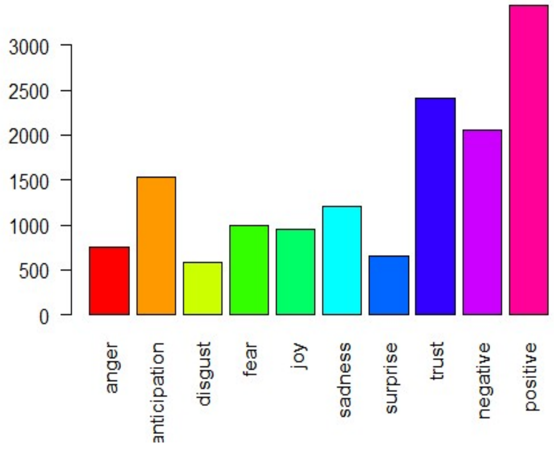

Figure 4 Sentiment Score/ Frequency

Figure 4 is obtained using the get_nrc_sentiment function of the syuzhet package. It gives us the score of various sentiments like anger, fear, joy, trust, anticipation, etc.

We plot this data using barplot and view the sentiments scores.

VII. Community Detection

The concept of community detection has emerged in network science as a method for finding groups within complex systems through represented on a graph. In contrast to more traditional decomposition methods which seek a strict block diagonal or block triangular structure, community detection methods find subnetworks with statistically significantly more links between nodes in the same group than nodes in different groups.

Node Degree: The degree of a node is the number of edges connected to the node. In terms of the adjacency matrix A, the degree for a node indexed by i in an undirected network is ki=∑jaij, where the sum is over all nodes in the network. In a directed network, each node has two degrees.

We plot Frequency of nodes against node degrees using histogram.

We use 'igraph' package for community detection: Routines for simple graphs and network analysis. It can handle large graphs very well and provides functions for generating random and regular graphs, graph visualization, centrality methods and much more.

Sociogram: A sociogram is a graphic representation of social links that a person has. It is a graph drawing that plots the structure of interpersonal relations in a group situation.

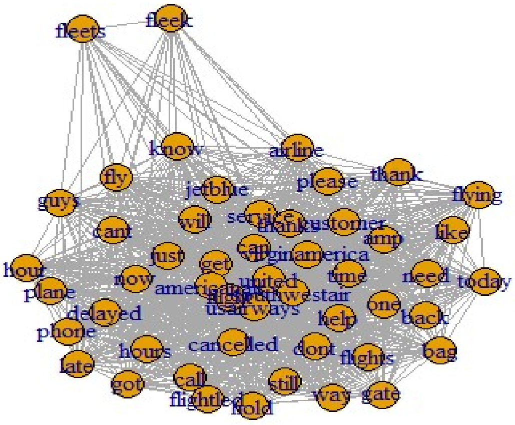

Figure 5 Sociogram

Edge Betweenness: The edge betweenness centrality is defined as the number of the shortest paths that go through an edge in a graph or network (Girvan and Newman 2002). Each edge in the network can be associated with an edge betweenness centrality value. An edge with a high edge betweenness centrality score represents a bridge-like connector between two parts of a network, and the removal of which may affect the communication between many pairs of nodes through the shortest paths between them.

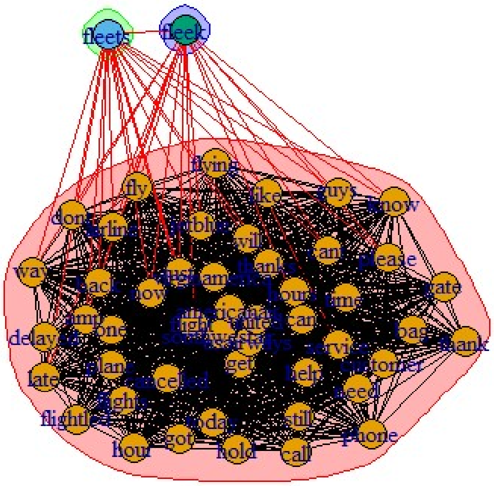

Figure 6 Edge Betweeness

Figure 6, gives the result for cluster_edge_betweenness performs this algorithm by calculating the edge betweenness of the graph, removing the edge with the highest edge betweenness score, then recalculating edge betweenness of the edges and again removing the one with the highest score, etc.

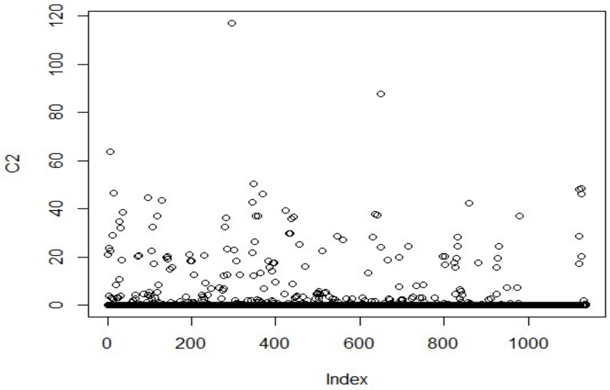

Figure 7 Estimate Edge Betweenness

Figure 7, gives us the result to estimate_betweenness, it only considers paths of length cutoff or smaller, this can be run for larger graphs, as the running time is not quadratic (if cutoff is small). If cutoff is zero or negative then the function calculates the exact betweenness scores.

Label propagation: is a heuristic method initially proposed for community detection in networks, while the method can be adopted also for other types of network clustering and partitioning. Among all the approaches and techniques described in this book, label propagation is neither the most accurate nor the most robust method. It is, however, without doubt one of the simplest and fastest clustering methods. Label propagation can be implemented with a few lines of programming code and applied to networks with hundreds of millions of nodes and edges on a standard computer, which is true only for a handful of other methods in the literature.

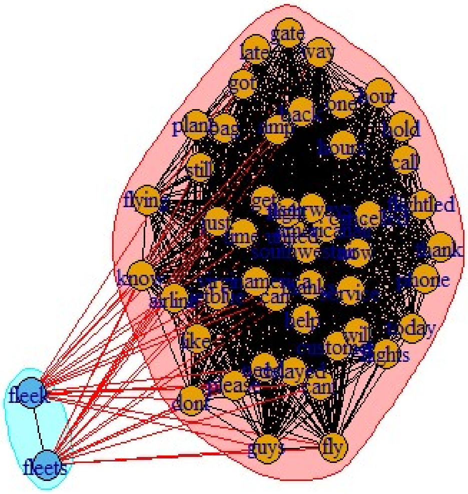

Figure 8 Propogative Labels

Figure 8, Gives us the result of Community Detection using Propogative Labels.

This is a fast, nearly linear time algorithm for detecting community structure in networks. In works by labeling the vertices with unique labels and then updating the labels by majority voting in the neighborhood of the vertex.

Greedy Optimisation:

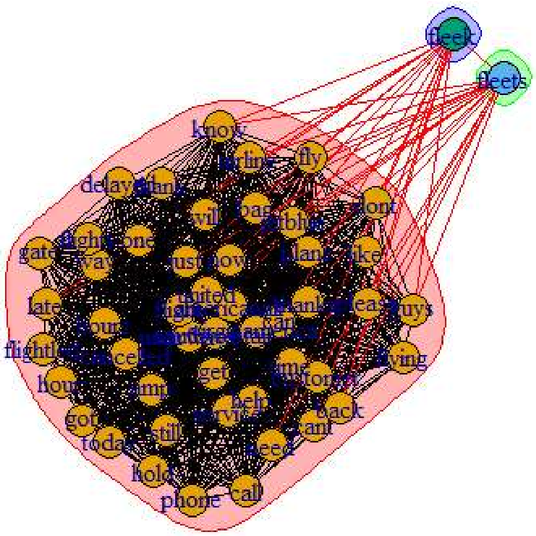

Figure 9 Greedy Optimization

Figure 9, This function implements the fast greedy modularity optimization algorithm for finding community structure. This function tries to find dense subgraph, also called communities in graphs via directly optimizing a modularity score.

VIII. Clustering

K Means Clustering: is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centres or cluster centroid), serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances.

Hierarchical Clustering: method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data points as a separate cluster. Then, it repeatedly executes the subsequent steps:

Identify the 2 clusters which can be closest together, and Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.

In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or cluster are break up (top-down view).

IX. Analysis and Inferences

- The more frequent a specific word appears in a source of text data the bigger and bolder it appears in the word cloud.
- The sentimental score tells us about the emotions and feelings from the text data.
- We can know how the people were during that situation with the help of the sentimental scores.
- The network diagram depicts the relation between the nodes.
- It is a labelled and undirected graph.
- The nodes(words) are connected with the help of distance vector.
- From the bar plots and data visualisation we can observe:
    - No. of tweets for different sentiments
    - Sentiment according to airlines
    - Count of Negative Reasons
    - Mean Air Sentiment Confidence
    - Mean Negative Reasons Confidence
- Edge betweenness determines the weakest links and gives us the removed links.
- We also observe from community detection algorithms that words like ‘fly’, ‘late’, ‘plan’, etc. are part of same network.

X. References

- *Arasteh, M., Alizadeh, S. A fast divisive community detection algorithm based on* *edge degree betweenness centrality. https://doi.org/10.1007/s10489-018-1297-9*

- *Handbook of Natural Language Processing, Second Edition, (editors: N. Indurkhya and F. J. Damerau), 2010*
- *Sathiyakumari K., Vijaya M.S. (2016) Community Detection Based on Girvan Newman Algorithm and Link Analysis of Social Media.*
- *WIKIPEDIA.ORG*
- *https://igraph.org/*
- *https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf*
- *https://www.red-gate.com/simple-talk/sql/bi/text-mining-and-sentiment-analysis-with-r/*
- *https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html*
- *https://www.displayr.com/what-is-hierarchical-clustering/#:~:text=Hierarchical%20clustering%2C%20also%20known%20as,broadly%20similar%20to%20each%20other*
- *https://developer.twitter.com/en/docs/tutorials/how-to-analyze-the-sentiment-of-your-own-tweets*
- *https://ieeexplore.ieee.org/document/8377739*
- *http://archive.ics.uci.edu/ml/datasets.php*
- *https://snap.stanford.edu/data/*