**CHAPTER 1**

**INTRODUCTION**

**1.1 Preamble**

Social media sites such as facebook, twitter provide great venues for people to share their experiences, complaints, stress and seek social support. On various social media sites, people discuss and share their everyday encounters in an informal and casual manner. People’s digital footprints provide vast amount of implicit knowledge and a whole new perspective for researchers to understand people’s experiences outside the controlled railway environment. This understanding can inform institutional decision-making on interventions for at-risk people for improvement of railway services. The abundance of social media data provides opportunities to understand people’s experiences and also raises methodological difficulties in making sense of social media data for its purposes.

Data mining has recently become one of the most progressive and promising fields for the extraction and manipulation of data to produce useful information. Thousands of businesses are using data mining applications every day in order to manipulate, identify, and extract useful information from the records stored in their databases, data repositories, and data warehouses.

Data mining which is also known as knowledge discovery is the process in which we extract useful information from the large set of the data. Now a days daily a huge amount of data in generated, a survey says that 90% of all the end the word is generated in past few years.  If we talk about the big data most of the data generated daily is in the form of unstructured data.

We are living in the data age where in every place you can see the data generation, if you are standing in queue for making reservation in the train at that place a large amount of data is generated continuously. Business society, medical field, Indian Railways, science and engineering and every aspect of life is generating a large amount of data daily.

**1.2 Need of the Project**

The emerging field of data analytics have focussed on analysing structured data obtained from Great Indian Railway, Ministry of Railway and some other pages from social sites like twitter and facebook. Data coming from these sites is unorganized and not much can be concluded from this data. The project is developed to analyse the data which is generated by tweets and categorize the complaints. These particular complaints are forwarded to the respective twitter pages of the railway department which is responsible to answer to these complaints. The passengers were not able to direct their complaints directly to the particular person responsible. If even, the authority was able to realize the complaints ,they were unable to categorize them as frequent complaints and complaints which needs immediate actions. The project gives a platform for twitter users, who have travelled in train and faced any kind of problem ,to directly forward their complaints to the authority ,so the user can be addressed directly. The project is developed to understand passengers’ problems and to forward them to particular railway department so that they can solve their problems easily and efficiently.

**1.3 Problem Statement**

Railway passengers have been struggling with many problems for a long time. They share their problem on social networks. Some of them are listed that have been frequently shared and are addressed in this:-

* Long queue in reservation system and less number of reservation counter.
* Frequent change of platform
* Heavy crowd in platform
* Problems of theft
* Disturbance due to sales person
* Unauthorised entry in the reserved compartment
* Bulk booking by agent.

The above mentioned problems and many other problems are shared by people on social networks such as twitter. Today ,there is no platform for these complaints to be directly forwarded to the respective department.Also,there is no procedure which analyses this complaints and categorises them. This research have categorised this complaints as:

* To be Immediately forwarded
* Frequently addressed complaints

All the “immediate” complaints are directly sent to the respective authority’s twitter page.All the other complaints are forwarded to the official government page of railway.

By analysing the complaints,the frequently addressed complaints it could be implied that there is some lack of rules and regulations that are addressing these issues in that sector. This system will suggest that there is requirement to make amendements in the rules and regulations that are not implemented effectively regarding the complaints that are arriving frequently.The suggestions of amendments of rules and regulations would be sent to the authority of Indian Railway .

**1.4 OBJECTIVE**

Reducing the barriers in the development of the indian Railways towards the passsengers problems, policies to be improved, and the judgement of frequent and immediate complaints and making automated system addressing the real time problem of the passengers of Indian Railways.

**1.5 SOLUTION APPROACH**

The solution approach can be defined as:

* Fetching the data from social networking sites of Indian railway such as twitter and facebook from their respective API’s. For data fetching there is need of to automate the API’s in program. This can be done using java, R, flumes etc. Java has provided tm package to facilitate this feature. R is oriented to data mining.
* After data fetching, the main task is to transformation of fetched data into required file format. Data transformation is one of the crucial tasks. The data transformation totally depends on the tool that is going to be used in the project. There are already tools available for this purpose.
* Then data required to be preprocessed. The preprocessing involves four steps:

1. Data cleaning:Data cleaning or data scrubbing is the process of detecting or correcting corrupt from a recorded set, table or database.
2. Data transformation: It is process of converting the data fetched in the given file format into standard file format.
3. Data reduction: It is the process of reducing the obtained data into meaningful and fruitful part of the data.
4. Discretization: Discretization is the process of putting values into buckets so that there are a limited number of possible states. The buckets themselves are treated as ordered and discrete values.

* The preprocessing can be done through tools like GATE, Weka.
* To extract information from complaints we have to perform text mining. The text mining refers to process of extracting high-quality data from given text. The first basic step of text mining in information retrieval. Information retrieval includes the extraction of abstract information from text. There are several algorithms available for information extraction such as supervised learning process to extract information from unstructured data that the system is going to deal. Then the text will be converted into text corpus by lexical analysis of the text data.
* After information retrieval, the next task is pattern recognition. Pattern recognition phase involves the recognition of patterns in the text. This can be done through supervised and unsupervised learning. The system is going to used unsupervised learning in which there are set of predefined classes or clusters.
* Tagging has to be done after pattern recognition. The system is going to use knowledge tag. Knowledge tag is a meta-information that describes some aspect of informational resource.
* Then we have to perform association analysis for hierarchical clustering.
* Then we have to analyse the result obtained from the clusters. Now the complaints have been classified into two subcategories: frequent and immediate. The cluster containing the immediate complain will be forwarded to the head of the Indian railway. The immediate complaints include health issue of the passenger.

The frequently addressed complains are going to be evaluated. These evaluations of frequent complaints will help the system to determine the lack in the policies of the Indian railway. The frequency of the complaints concerned to the particular issue will be counted. This frequency of complaints will help us to determine that, that kind of complaints are arriving very frequently but there is some lapses in policy that’s why the same kind of complain have higher frequency.

**1.6 ORGANIZATION OF THE PROJECT REPORT**

The organization of rest of the project report is as follows:

Chapter 2 explains the hardware and software requirements, tools and technologies used for the project. The analysis part of our project which includes various functional as well as non-functional requirement of our project is dealt in Chapter 3. Chapter 4 gives the design and architectural details of our project and also include UML diagrams. The implementation details like hardware and softwares used to make the project, classes and methods used and the algorithms implemented are mentioned in Chapter 5. Chapter 6 is testing and results in which various scenarios in the project were tested and the corresponding outputs are shown. Chapter 7 concludes the project and gives the future work to be carried out.

**CHAPTER 2**

**BACKGROUNG STUDY AND FUNDAMENTALS**

This chapter explains the hardware and software requirements, tools and technologies used for the project.

**2.1 Software requirements**

* **twitteR Package**

twitteR is an R package which 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. The Twitter R package is a great way to get started in Text Analytics. This package is about how to get setup to pull twitter data into R so that you can do some text analytics with it. Basically it is only the first step of getting data in a text analytics Workflow.

* **StreamR Package**

StreamR Package allows to do real time analytics on data streams. This can be very usefull if you are working with large datasets which are already hard to put in RAM completely, let alone to build some statistical model on it without getting into RAM problems.

Most of the standard statistical algorithms require access to all data points and make several iterations over the data and are less suited for usage in R on big datasets. The stream package is currently focussed on clustering algorithms. The stream package allows you to **easily extend the use of the models with different data sources.** These can be SQL sources, Hadoop, Storm, Hive, simple csv files, flat files or other connections. It is quite easy to extend it towards other connections.

Streaming algorithms on the other hand are characterised by

* + single passes over the data,
  + using a limited amount of storage space and RAM
  + work in a limited amount of time
  + be ready to use the model at any time

* **Factoextra**

Provides some easy-to-use functions to extract and visualize the output of multivariate data analyses, including 'PCA' (Principal Component Analysis), 'CA' (Correspondence Analysis), 'MCA' (Multiple Correspondence Analysis), 'FAMD' (Factor Analysis of Mixed Data), 'MFA' (Multiple Factor Analysis) and 'HMFA' (Hierarchical Multiple Factor Analysis) functions from different R packages. It contains also functions for simplifying some clustering analysis steps and provides 'ggplot2' - based elegant data visualization. Principal component analysis (PCA) reduces the dimensionality of multivariate data, to two or three that can be visualized graphically with minimal loss of information. fviz\_pca() provides ggplot2-based elegant visualization of PCA outputs from: i) prcomp and princomp [in built-in R stats], ii) PCA [in FactoMineR], iii) dudi.pca [in ade4] and epPCA [ExPosition]

* **Ggplot2 package**

A system for 'declaratively' creating graphics, based on ``The Grammar of Graphics''. You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.

* **Rjava**

**rJava** is a simple R-to-Java interface. It is comparable to the .C/.Call C interface. rJava provides a low-level bridge between R and Java (via JNI). It allows to create objects, call methods and access fields of Java objects from R. This provides detailed guidance on interfacing R to Java archives inside an R package. The objective is to help other people to make available Java algorithms to the R world, be it to compare results, or for their own sake.

* **Plyr**

A set of tools that solves a common set of problems: you need to break a big problem down into manageable pieces, operate on each piece and then put all the pieces back together. For example, you might want to fit a model to each spatial location or time point in your study, summarise data by panels or collapse high-dimensional arrays to simpler summary statistics. The development of 'plyr' has been generously supported by 'Becton Dickinson'.

**Tidytext**

Using [tidy data principles](https://www.jstatsoft.org/article/view/v059i10) can make many text mining tasks easier, more effective, and consistent with tools already in wide use.  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. Much of the infrastructure needed for text mining with tidy data frames already exists in packages like dplyr, broom, tidyr and ggplot2.

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.

**RTextTools**

RTextTools has largely been used for topic classification in the social sciences. However, recent discussions with researchers at various universities have demonstrated that the package can be applied to a host of problems in the natural sciences as well.  
One such application is using text classification to identify breast cancer masses as benign or malignant. Using the Wisconsin Diagnostic Breast Cancer Dataset from UC Irvine, we wrote a script that trains eight classifiers on characteristics such as clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. When run on the data, the classifiers were able to achieve up to 96% recall accuracy on a randomly sampled training set of 200 patients and test set of 400 patients.

**CLUTO**

CLUTO is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters. CLUTO is well-suited for clustering data sets arising in many diverse application areas including information retrieval, customer purchasing transactions, web, GIS, science, and biology. CLUTO's distribution consists of both stand-alone programs and a library via which an application program can access directly the various clustering and analysis algorithms implemented in CLUTO.

**gCLUTO**

gCLUTO is a cross-platform graphical application for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters. gCLUTO is build on-top of the CLUTO clustering library. gCLUTO provides tools for visualizing the resulting clustering solutions using tree, matrix, and an OpenGL-based mountain visualization.

**SnowballC**

**Snowball stemmers are based on the C libstemmer UTF-8 library**

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. This dynamically determines the names of the languages for which stemming is currently supported by this package. The language names in lower case are returned, though please note that two- and three- letter ISO- 639 codes are also accepted by wordStem. This queries the C code for the list of languages that were compiled when the package was installed which in turn is determined by the code that was included in the distributed package itself.

**Tm package**

A framework for text mining applications within R. This dataset holds 50 news articles with additional meta information from the Reuters-21578 data set. All documents belong to the topic acq dealing with corporate acquisitions. The tm package offers functionality for managing text documents, abstracts the process of document manipulation and eases the usage of heterogeneous text formats in R. The package has integrated database back-end support to minimize memory demands. An advanced meta data management is implemented for collections of text documents to alleviate the usage of large and with meta data enriched document sets.[1]

**2.2 Tools and Technologies used**

**Rstudio**

RStudio is a [free and open-source](https://en.wikipedia.org/wiki/Free_and_open-source) [integrated development environment](https://en.wikipedia.org/wiki/Integrated_development_environment) (IDE) for [R](https://en.wikipedia.org/wiki/R_(programming_language)), a [programming language](https://en.wikipedia.org/wiki/Programming_language) for [statistical computing](https://en.wikipedia.org/wiki/Statistical_computing) and graphics. RStudio is available in two editions: RStudio Desktop, where the program is run locally as a regular [desktop application](https://en.wikipedia.org/wiki/Desktop_application); and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote [Linux](https://en.wikipedia.org/wiki/Linux) server. Prepackaged distributions of RStudio Desktop are available for [Windows](https://en.wikipedia.org/wiki/Windows), [macOS](https://en.wikipedia.org/wiki/MacOS), and Linux. RStudio is written in the [C++](https://en.wikipedia.org/wiki/C%2B%2B) programming language.[2]

**R language**

R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering) and graphical techniques. R is available as Free Software under the terms of the [Free Software Foundation](http://www.gnu.org/)’s [GNU General Public License](https://www.r-project.org/COPYING) in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS. R is an integrated suite of software facilities for data manipulation, calculation and graphical display[3]. It includes

* an effective data handling and storage facility,
* a suite of operators for calculations on arrays, in particular matrices,
* a large, coherent, integrated collection of intermediate tools for data analysis,
* graphical facilities for data analysis and display either on-screen or on hardcopy, and
* a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

**2.3 Hardware Requirements**

Processor - AMD A10, i3,i5,i7

RAM - 4 GB

Hard drive - 512GB

Operating system : Windows 7 and above, Linux

**CHAPTER 3**

**ANALYSIS**

The analysis part of our project includes various functional as well as non-functional requirement of our project is dealt in this chapter. This phase was initialized by manually analyzing the complaints in form of tweets of the Indian railway twitter page.

**3.1 Detailed Problem Statement**

Indian Railways is the backbone of the country.The Railway System of any country is not completely perfect and as said, the same is true for Indian Railways. Railway passengers have been struggling with many problems for a long time. There are many kind of problems experienced by people while travelling in trains and even while boarding the train like problems in booking tickets, reservation ambiguities, problems related with theft. People have a need to find a platform through which they share their complaints directly with the persons who are answerable for complaints in such areas .At present they are not having any software through which they can be directly connected to the railway authorities.

Therefore, People generally share their problems with the people they know and many of times ,they also post these problems on social media. They share their problem on social networks like facebook, twitter etc. If these problems are addressed by the persons who have a say in the solutions for the problems then it can help the passengers to get a valid action regarding their problem as soon as possible.These problems ,if taken care of may change the present condition of railways in our country.Therefore,we are trying to provide a platform to connect passengers’ complaints directly with the officials who are answerable for such complaints. Some of them are listed that have been frequently shared and are addressed in this:-

* Long queue in reservation system and less number of reservation counter.
* Frequent change of platform
* Heavy crowd in platform
* Problems of theft
* Disturbance due to sales person
* Unauthorised entry in the reserved compartment
* Bulk booking by agent.

The above mentioned problems and many other problems are shared by people on social networks such as twitter.Today,there is no platform for these complaints to be directly forwarded to the respective department. Also, there is no procedure which analyses this complaints and categorises them. This research have categorised this complaints as the complaints which need to be Immediately forwarded or Frequently addressed complaints. These could be done by keeping log of complaints which were registered earlier by the system.

All the “immediate” complaints are directly sent to the respective authority’s twitter page. All the other complaints are forwarded to the official government page of railway.These would be done by clustering the keywords which would contribute to the field which needs some kind of immediate actions. Also the complaints which are arriving on a regular basis suggests that they are not being addressed accurately and appropriately.The system would suggest this problem via a tweet directly to the railway department of India.

By analysing the complaints,the frequently addressed complaints it could be implied that there is some lack of rules and regulations that are addressing these issues in that sector. This system will suggest that there is requirement to make amendements in the rules and regulations that are not implemented effectively regarding the complaints that are arriving frequently.The suggestion of amendments in rules and regulations would be sent to the authority of Indian Railway .Thus the system would bridge the gap between an ordinary passenger having complaints and the railway authorities of India.

**3.2 Requirement Analysis**

Requirement is a condition of capability of a system required by customer to solve a problem or achieve an objective. It is a capability that a product must poses in order to ultimately satisfy the customer, contract, standards, specification or other formally imposed documents. It is a document representation of condition or capabilities.

Requirements analysis, also called requirements engineering, is the process of determining user expectations for a new or modified product. These features, called requirements, must be quantifiable, relevant and detailed.

The project of Data mining for Railway Complaints based on the service where network will be used, and it is based on functional and non functional requirements. The functional requirements are-

***3.2.1 Functional Requirements***

Functional requirement defines a function of a [system](https://en.wikipedia.org/wiki/System) or its component. Functional requirements are the behaviour that the system must support. These are the attributes that characterise what the software does to fulfil the need of the customer.

A major vehicle for describing functional requirements are use cases and UML use case diagrams.

**Fetch data** : Data is fetched from social networking sites such as facebook and twitter from their API’s. Data will be in the form of valid textual string.

**Preprocess** : The fetched data is preprocessed which means that data which is fetched in given file format is converted into standard file format.

**Clustering**: The Preprocessed data will be classified based on the clusters. Clustering will be done with the help of data mining tools. For our project we are using WEKA. Clustering will be done on the basis of some keywords that we will define.

**Evaluation**: The data will be evaluated and categorised according to the type of complain. There will be three categories- immediate, frequent and others complaint.

**Forwarding**: All the “immediate” complaints are directly sent to the respective authority’s twitter page.All the other complaints are forwarded to the official government page of railway. By analysing the complaints,the frequently addressed complaints it could be implied that there is some lack of rules and regulations that are addressing these issues in that sector. This system will suggest that there is requirement to make amendements in the rules and regulations that are not implemented effectively regarding the complaints that are arriving frequently.The suggestions of amendments of rules and regulations would be sent to the authority of Indian Railway .

***3.2.2 Non Functional Requirements***

      The non-functional requirements in data mining could come from the operating environment, the users, and the competitive products. In the operating environment, data can be affected by the system which is used in supporting the process. It poses problem on how the software will work towards establishing dynamic data architecture. Furthermore, users are also behind the non-functional requirements for data mining in software engineering primarily because they control a big fraction of the entire program and they are the ones who completely understand the attributes of the system. Lastly, the existence of competitive alternatives affects the non-functional requirements because of their features which generally affect the quality of the system.

**Fast Delivery**: Our system will take care of delivery of complaints so that the complaints can be delivered fast and their issues can be resolved.

**Response Time:** It is appropriate for real-time and because of categorising the complaints in immediate and frequent categories, it will be easy for Railway Minister to response to user’s complaints.

**Adaptability:** It is adaptable for further improvements according to user’s and organisation’s need.

**Security:** The system will be made such that the data is secured.

**Resilient:** It will be able to provide and maintain an acceptable level of service so that the users can be satisfied.

**3.3 FEASIBILITY STUDY**

**1. Technical Feasibility** - It includes technical resources to be handle and the technology is in the market or not. In this project we are combining the Indian railways, which is a backbone of our nation and the twitter social media, the concept of the mining the data of Indian Railways is new, which is fruitful to the nation, to which department indian Railways are lacking. Also here we are using many tools for egweka for the testing purpose which generates the graph and an environment for the data mining to be perform on the datset. The dataset in our project is the twitter api, which takes the feasible amount of time.

2.**Economic:** It contains money aspect of the project. In this project

**3.Operational feasibilty**: After the technical and economical aspect we study the aperationalfeasibilty , that is the insertion of new system with the existng system, Here we are connecting our system eith twitter api and indian railways site , which is on internet,and being the www, the system will not have any problem of being compatible and also the tool will result the dataset we required.

**4.Scheduling :** Time of completion of proejct .This project will be completed by the may-june 2017.

**5. Legal and Startegic** : As indianarailways are government organisation and due to right to know about the government organisation and using the open source software no legal proceeding will be there.

**3.4 DIAGRAMS**

**3.4.1 Flowchart**

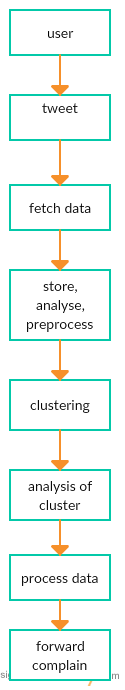
****

Figure 3.1 - Flowchart

**3.4.2 Use Case Diagram**

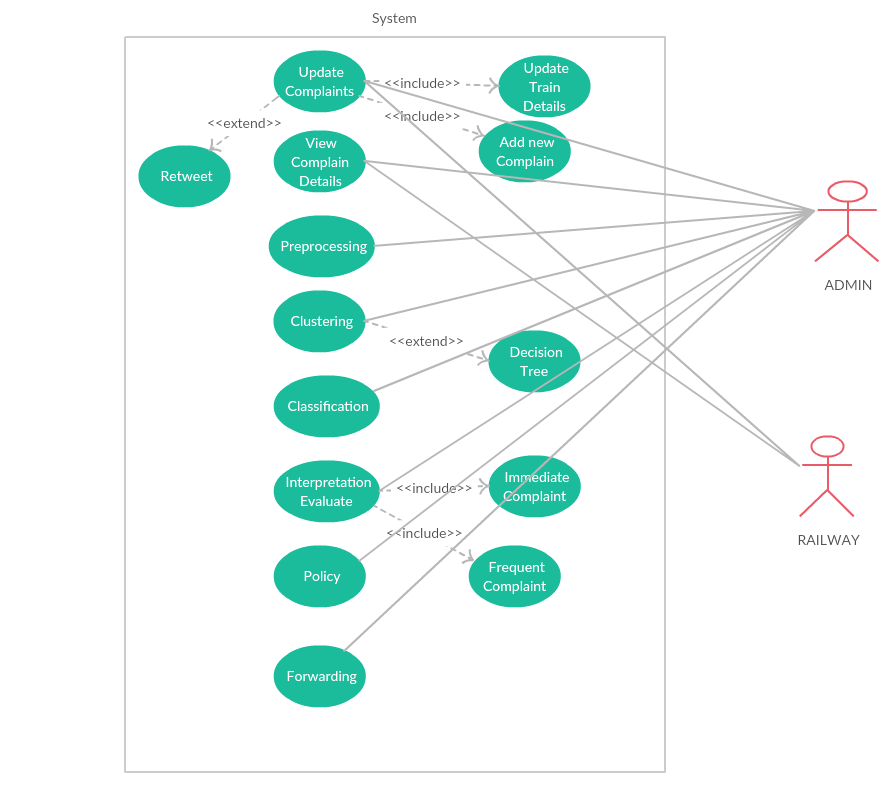
****

Figure 3.2 - Use Case Diagram

**Use Case Description**

**Update Complaint-**

Pre Condition: User should have a twitter account.

Flow of events: User tweets a valid complaint.

Post Condition: Complaint is updated on the Indian Railways’ twitter page.

**Preprocessing**-

Pre Condition: Run time dynamic data should be fetched by the system.

Flow of events: Run time dynamic data is processed by the software and redundancy and noise are removed.

Post Condition: The data should be now ready to be fed into clustering.

**Classification**-

Pre Condition: The classes shoud be properly defined and trained by the training data set.

Flow of events: The clustered data goes into specified classes.

Post Condition: Classification face of data is completed.

**Forwarding-**

Pre Condition: All the pre requisite like email ids, should be available.

Flow of events: The categorized data is sent via mail to the particular Railway department.

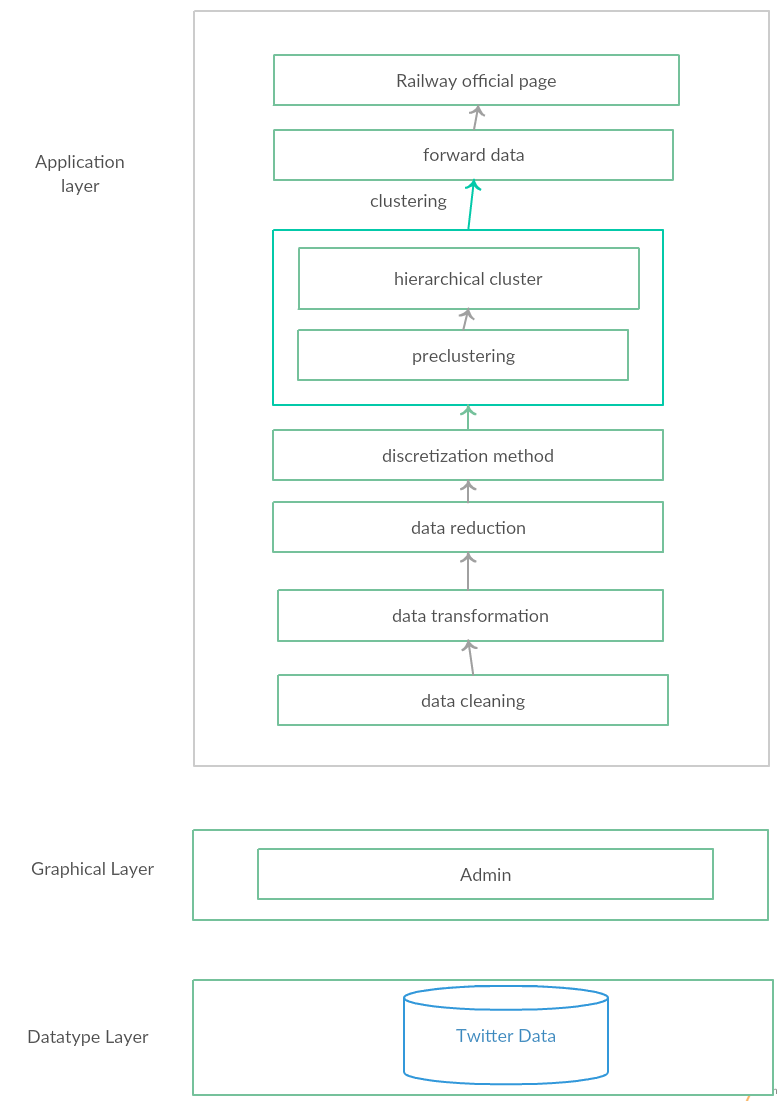
Post Condition: The mail is sent to the head of particular Railway departments.

**CHAPTER 4**

**DESIGN**

This chapter gives the design and architectural details of our project. This chapter also includes UML diagrams like activity diagrams, state chart diagrams, class diagram and component diagram which describes the system well.

**4.1 Architectural Diagram**

****

**Figure 4.1 - Architectural Diagram**

**4.2 E-R Diagram**

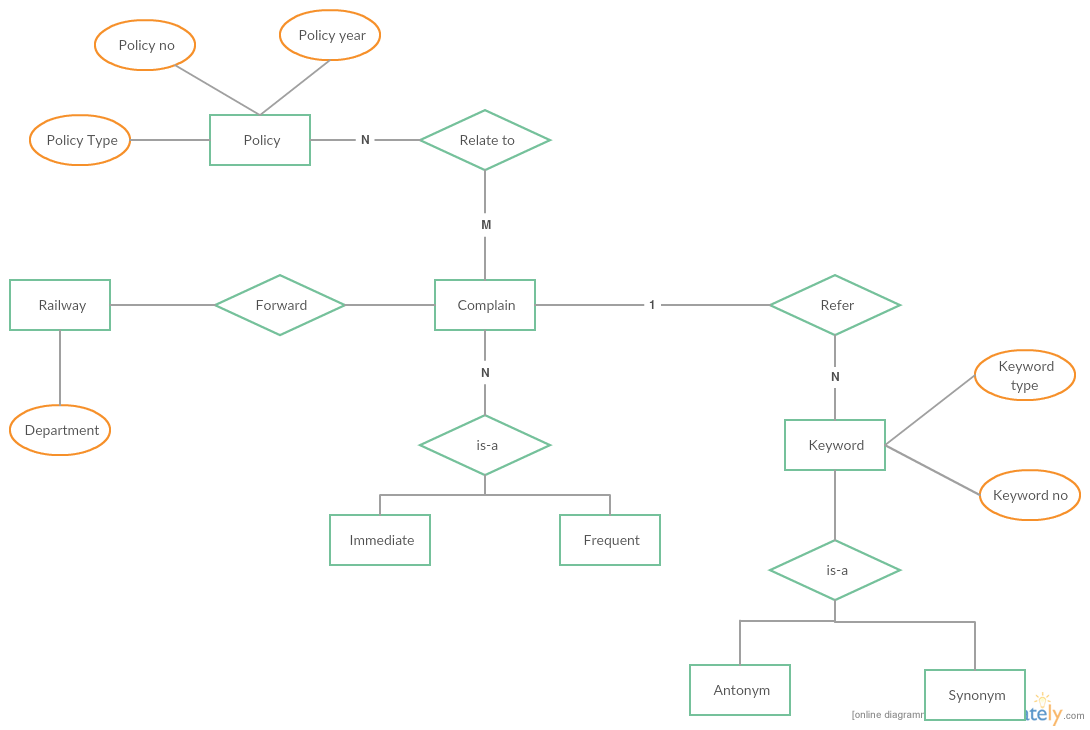


Figure 4.2 - E-R Diagram

**4.3 Activity Diagram**

Activity diagrams are graphical representations of workflow of stepwise activities and actions with support for choice, iteration and concurrency. In the unified modelling language, activity diagrams are intended to model both computational and organizational processes (i.e. workflows). Activity diagrams show the overall flow of control.

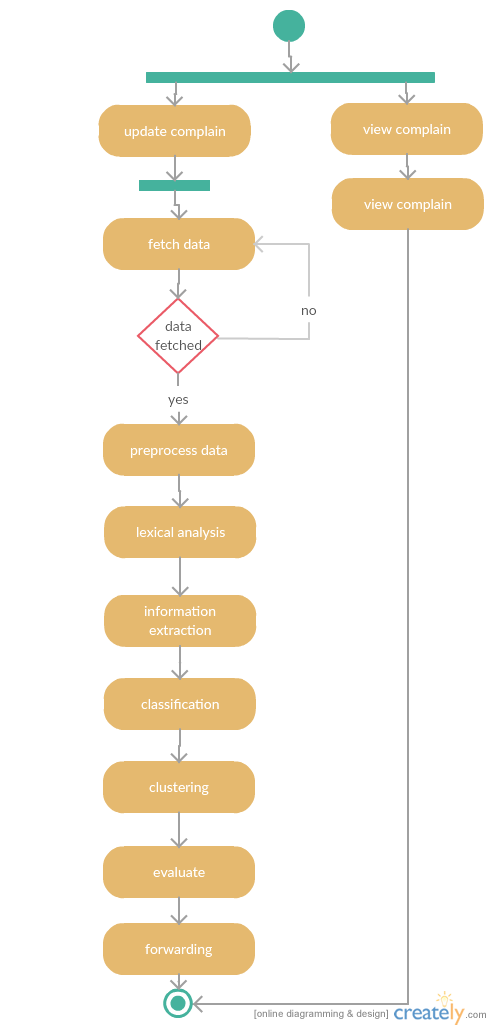


Figure 4.3 - Activity Diagram

4.3.1 ACTIVITY DIAGRAM FOR UPDATE COMPLAINTS

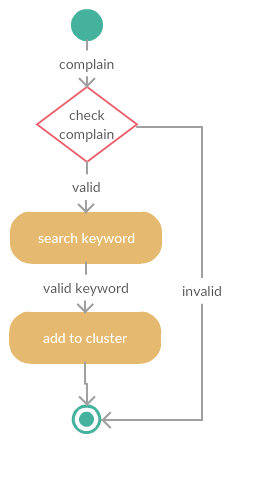


Figure 4.4 – Activity Diagram for Update Complaints

4.3.2 ACTIVITY DIAGRAM FOR FORWARDING

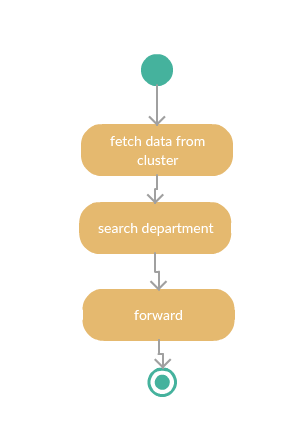


Figure 4.4 - Activity Diagram for Forwarding

4.3.3 ACTIVITY DIAGRAM FOR CLUSTERING

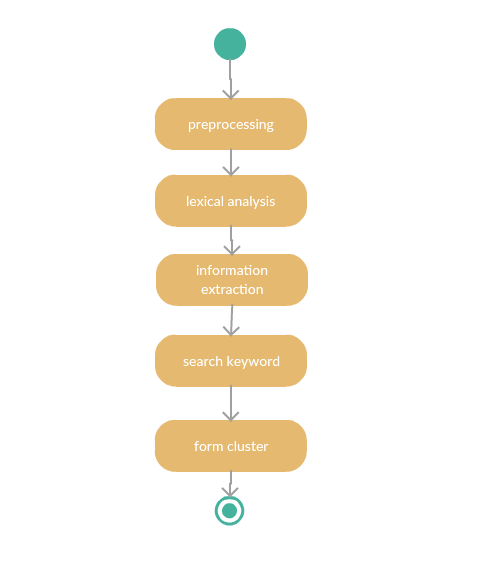


Figure 4.5 – Activity Diagram for Clustering

..

**4.4 State Chart Diagram**

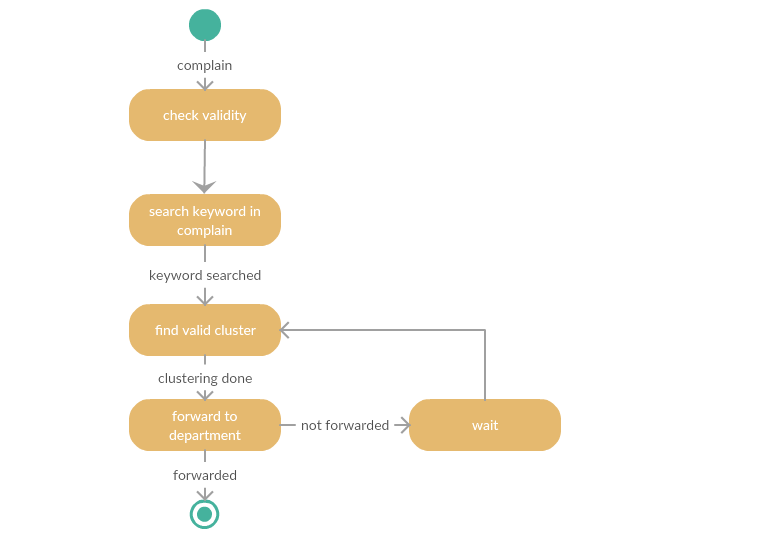


Figure 4.6 - State Chart Diagram For Update Complaints

4.4.1 STATE CHART DIAGRAM FOR VIEW COMPLAINTS DETAILS

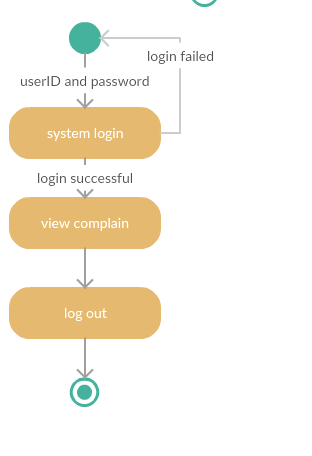


Figure 4.7- State Chart Diagram For View Complaint Details

4.4.2 STATE CHART DIAGAM FOR CLUSTERING

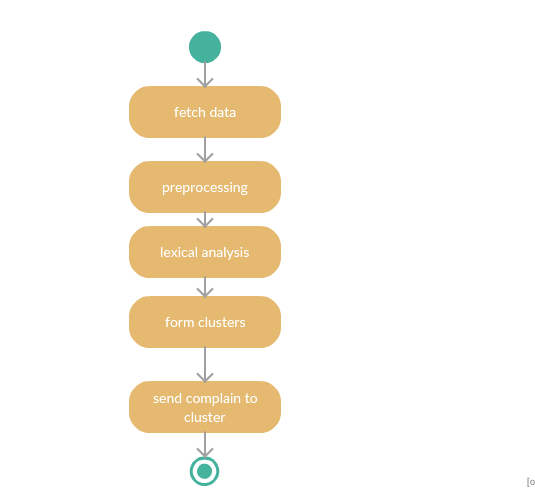


Figure 4.8 - State Chart Diagram for Clustering

4.4.2 STATE CHART DIAGRAM FOR FORWARDING

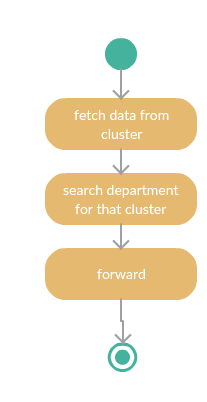


Figure 4.9 - State Chart Diagram for Forwarding

**4.5 Class Diagram**

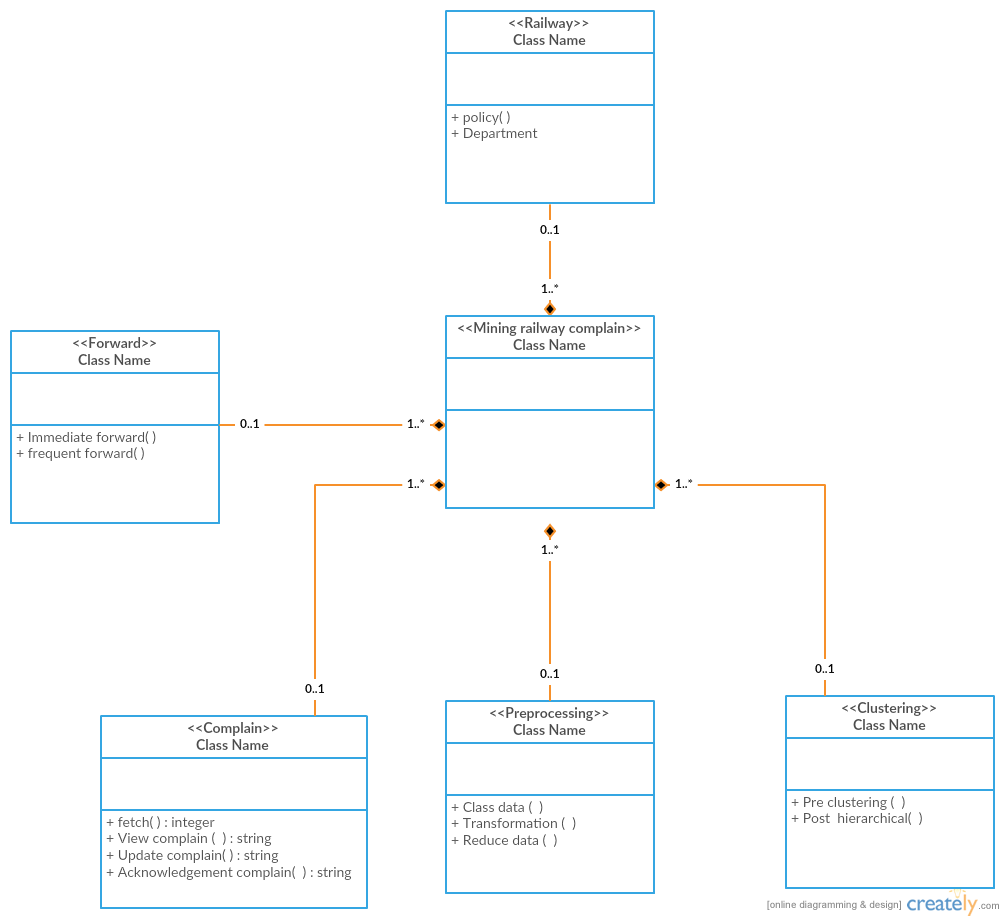
****

Figure 4.10: Class Diagram

**4.6 Component Diagram**

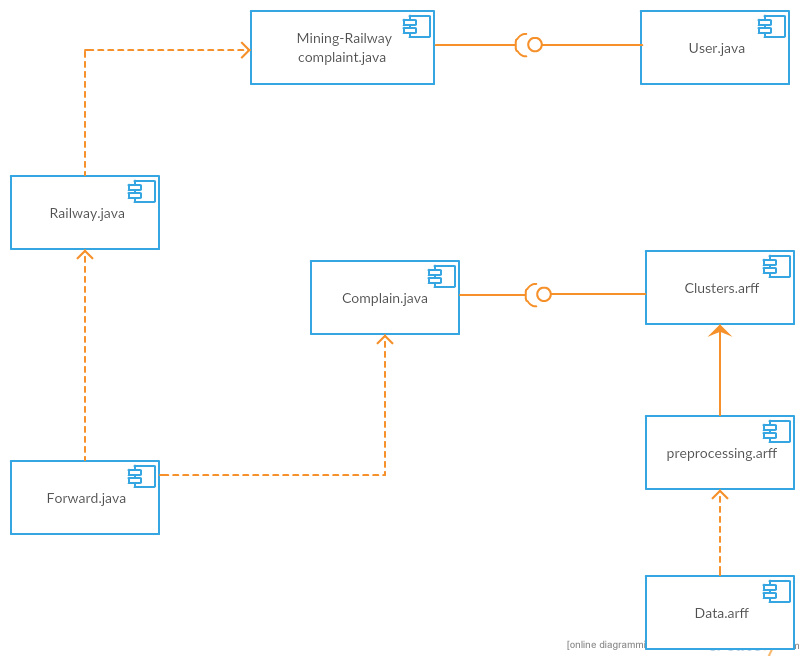


Figure 4.11: Component Diagram

**CHAPTER 5**

**IMPLEMENTATION**

Implementation is the realization of a technical specification or algorithm as a program, software component, or other computer system through computer programming and development. Many implementations may exist for a given specification or standard**.**

**5.1 SUPPORT VECTOR MACHINE (SVM) ALGORITHM**

**Preliminaries:**

Linear Classifiers Support vector machines are an example of a linear two-class classifier. This section explains what that means. The data for a two class learning problem consists of objects labeled with one of two labels corresponding to the two classes; for convenience we assume the labels are +1 (positive examples) or -1 (negative examples). In what follows boldface x denotes a vector with components xi. The notation xi will denote the ith vector in a dataset f(xi; yi)gni*=1*, where yi is the label associated with xi. The objects xi are called patterns or examples. We assume the examples belong to some set X . Initially we assume the examples are vectors, but once we introduce kernels this assumption will be relaxed, at which point they could be any continuous/discrete object (e.g. a protein/DNA sequence or protein structure)[5].

A key concept required for defining a linear classifier is the dot product between two vectors, also referred to as an inner product or scalar product, de need as wTx = Pi wixi. A linear classifier is based on a linear discriminant function of the form

|  |  |
| --- | --- |
| f(x) = wTx + b: | (1) |

The vector w is known as the weight vector, and b is called the bias. Consider the case b = 0 rst. The set of points x such that wTx = 0 are all points that are perpendicular to w and go through the origin | a line in two dimensions, a plane in three dimensions, and more generally, a hyperplane. The bias b translates the hyperplane away from the origin. The hyperplane

|  |  |
| --- | --- |
| fx : f(x) = wTx + b = 0g | (2) |

divides the space into two: the sign of the discriminant function f(x) denotes the side of the hyperplane a point is on (see 1). The boundary between regions classified as positive and negative is called the decision boundary of the classifier. The decision boundary de need by a hyperplane is said to be linear because it is linear in the input examples (c.f. Equation 1). A classifier with a linear decision boundary is called a linear classifier. Conversely, when the decision boundary of a classifier depends on the data in a non-linear way (see Figure 4 for example) the classifier is said to be non-linear.

**Large Margin Classification**

In what follows we use the term linearly separable to denote data for which there exists a linear decision boundary that separates positive from negative examples (see Figure 2). Initially we will assume linearly separable data, and later indicate how to handle data that is not linearly separable.

**The Geometric Margin**

In this section we de ne the notion of a margin. For a given hyper plane we denote by x*+*(x ) the closest point to the hyper plane among the positive (negative) examples. The norm of a vector w p denoted by ||w|| is its length, and is given by wTw. A unit vector w^ in the direction of w is given by w=||w|| and has ||w^|| = 1. From simple geometric considerations the margin of a hyper plane f with respect to a dataset D can be seen to be:

|  |  |
| --- | --- |
| MD(f) ~~=~~1/2\_wT(x+-x\_) |  |

where w^ is a unit vector in the direction of w, and we assume that x*+* and x are equidistant from the decision boundary i.e.

|  |  |  |  |
| --- | --- | --- | --- |
| f(x*+*) | = | wTx*+* + b = a |  |
| f(x ) | = | wTx + b = a | (8) |

for some constant a > 0. Note that multiplying our data points by a fixed number will increase the margin by the same amount, whereas in reality, the margin hasn't really changed | we just changed the \units" with which we measure it. To make the geometric margin meaningful we x the value of the decision function at the points closest to the hyper plane, and set a = 1 in Eqn. (8). Adding the two equations and dividing by ||w|| we obtain:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MD(f) ~~=~~1/2\_wT(x+ +x\_) |  |  |  |  |  |  |  |  |  |  |

**Support Vector Machines**

Now that we have the concept of a margin we can formulate the maximum margin classifier. We will rst de ne the hard margin SVM, applicable to a linearly separable dataset, and then modify it to handle non-separable data. The maximum margin classifier is the discriminant function that maximizes the geometric margin 1=||w|| which is equivalent to minimizing jjwjj*2*. This leads to the following constrained optimization problem:



The constraints in this formulation ensure that the maximum margin classifier each example correctly, which is possible since we assumed that the data is linearly separable. In practice, data is often not linearly separable; and even if it is, a greater margin can be achieved by allowing the classifier to misclassify some points. To allow errors we replace the inequality constraints in Eqn. (10) with

yi(wTxi + b) >1 -ξ i=1.....n

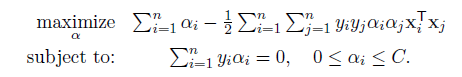
where i 0 are slack variables that allow an example to be in the margin (0 i 1, also called a margin error) or to be misclassified ( i > 1). Since an example is misclassified if the value of its P

slack variable is greater than 1, i i is a bound on the number of misclassified examples. Our objective of maximizing the margin, i.e. minimizing ( *12)* ||w||*2* will be augmented with a term C Pi i to penalize misclassification and margin errors. The optimization problem now becomes:

Minimize(w,b)  (1/2) || W ||2 + C ∑i=1 to n ξi

Subject to: Y(i) (wT X(i)+b) > 1 i=1.....n 1- ξi ξi>0

The constant C > 0 sets the relative importance of maximizing the margin and minimizing the amount of slack. This formulation is called the soft-margin SVM, and was introduced by Cortes and Vapnik [11]. Using the method of Lagrange multipliers, we can obtain the dual formulation which is expressed in terms of variables i [5]:



The dual formulation leads to an expansion of the weight vector in terms of the input examples:

W= ∑i=1 to n Yiαi Xi

The examples xi for which i > 0 are those points that are on the margin, or within the margin when a soft-margin SVM is used. These are the so-called support vectors. The expansion in terms of the support vectors is often sparse, and the level of sparsity (fraction of the data serving as support vectors) is an upper bound on the error rate of the classifier .The dual formulation of the SVM optimization problem depends on the data only through dot products. The dot product can therefore be replaced with a non-linear kernel function, there by performing large margin separation in the feature-space of the kernel (see Figures 4 and 5). The SVM optimization problem was traditionally solved in the dual formulation, and only recently it was shown that the primal formulation, Equation (11), can lead to e client kernel-based learning .

**Understanding the Effects of SVM and Kernel Parameters**

Training an SVM nds the large margin hyperplane, i.e. sets the parameters i and b (c.f. Equation 4). The SVM has another set of parameters called hyperparameters: The soft margin constant, C, and any parameters the kernel function may depend on (width of a Gaussian kernel or degree of a polynomial kernel). In this section we illustrate the effect of the hyperparameters on the decision boundary of an SVM using two-dimensional examples.

We begin our discussion of hyper parameters with the soft-margin constant, whose role is illustrated in Figure 3. For a large value of C a large penalty is assigned to errors/margin errors. This is seen in the left panel of Figure 3, where the two points closest to the hype rplane affect its orientation, resulting in a hyper plane that comes close to several other data points. When C is decreased (right panel of the gure), those points become margin errors; the hyper plane's orientation is changed, providing a much larger margin for the rest of the data Kernel parameters also have a significant effect on the decision boundary. The degree of the polynomial kernel and the width parameter of the Gaussian kernel control the flexibility of the resulting classifier (Figures 4 and 5). The lowest degree polynomial is the linear kernel, which is not sufficient when a non-linear relationship between features exists. For the data in Figure 4 a degree-2 polynomial is already flexible enough to discriminate between the two classes with a sizable margin. The degree-5 polynomial yields a similar decision boundary, albeit with greater curvature. Next we turn our attention to the Gaussian kernel: k(x; x0) = exp( ||x x0||*2*). This expression is essentially zero if the distance between x and x0 is much larger than 1=p ; i.e. for a fixed x0 it is localized to a region around x0. The support vector expansion, Equation (4) is thus a sum of Gaussian \bumps" centered around each support vector. When is small (top left panel in Figure 5) a given data point x has a non-zero kernel value relative to any example in the set of support vectors.

Therefore the whole set of support vectors affects the value of the discriminant function at x, resulting in a smooth decision boundary. As is increased the locality of the support vector expansion increases, leading to greater curvature of the decision boundary. When is large the value of the discriminant function is essentially constant outside the close proximity of the region where the data are concentrated (see bottom right panel in Figure 5). In this regime of the parameter the classifier is clearly over fitting the data. As seen from the examples in Figures 4 and 5 the parameter of the Gaussian kernel and the degree of polynomial kernel determine the flexibility of the resulting SVM in fitting the data. If this complexity parameter is too large, over fitting will occur (bottom panels in Figure 5).A question frequently posed by practitioners is \which kernel should I use for my data?" There are several answers to this question. The rst is that it is, like most practical questions in machine learning, data-dependent, so several kernels should be tried. That being said, we typically follow the following procedure: Try a linear kernel first, and then see if we can improve on its performance using a non-linear kernel. The linear kernel provides a useful baseline, and in many bioinformatics applications provides the best results.Furthermore, an SVM with a linear kernel is easier to tune since the only parameter that affects performance is the soft-margin constant. Once a result using a linear kernel is available it can serve as a baseline that you can try to improve upon using a non-linear kernel. Between the Gaussian and polynomial kernels, our experience shows that the Gaussian kernel usually outperforms the polynomial kernel in both accuracy and convergence time.

**Model Selection**

The dependence of the SVM decision boundary on the SVM hyper parameters translates into a dependence of classifier accuracy on the hyper parameters. When working with a linear classifier the only hyper parameter that needs to be tuned is the SVM soft-margin constant. For the polynomial and Gaussian kernels the search space is two-dimensional. The standard method of exploring this two dimensional space is via grid-search; the grid points are generally chosen on a logarithmic scale and classifier accuracy is estimated for each point on the grid. This is illustrated in Figure 6. A classifier is then trained using the hyper parameters that yield the best accuracy on the grid. The accuracy landscape in Figure 6 has an interesting property: there is a range of parameter values that yield optimal classifier performance; furthermore, these equivalent points in parameter space fall along a \ridge" in parameter space. This phenomenon can be understood as follows. Consider a particular value of ( ; C). If we decrease the value of , this decreases the curvature of the decision boundary; if we then increase the value of C the decision boundary is forced to curve to accommodate the larger penalty for errors/margin errors. This is illustrated in Figure 7 for two dimensional data.

**Normalization**

Large margin classifiers are known to be sensitive to the way features are scaled [14]. Therefore it is essential to normalize either the data or the kernel itself. This observation carries over to kernel-based classifiers that use non-linear kernel functions: The accuracy of an SVM can severely degrade if the data is not normalized . Some sources of data, e.g. microarray or mass-spectrometry data require 0normalization methods that are technology-specific. In what follows we only consider normalization methods that are applicable regardless of the method that generated the data.

Normalization can be performed at the level of the input features or at the level of the kernel (normalization in feature space). In many applications the available features are continuous values, where each feature is measured in a different scale and has a different range of possible values. In such cases it is often beneficial to scale all features to a common range, e.g. by standardizing the data (for each feature, subtracting its mean and dividing by its standard deviation). Standardization is not appropriate when the data is sparse since it destroys sparse since each feature will typically have a different normalization constant. Another way to handle features with different ranges is to bin each feature and replace it with indicator variables that indicate which bin it falls in. An alternative to normalizing each feature separately is to normalize each example to be a unit vector. If the data is explicitly represented as vectors you can normalize the data by dividing each vector by its norm such that ||x|| = 1 after normalization. Normalization can also be performed at the level of the kernel, i.e. normalizing in feature-space, leading to ||(x)|| = 1 (or equivalently k(x; x) = 1). This is accomplished using the cosine kernel which normalizes a kernel k(x; x0) to:

|  |  |  |
| --- | --- | --- |
|  |  |  |

Note that for the linear kernel cosine normalization is equivalent to division by the norm. The use of the cosine kernel is redundant for the Gaussian kernel since it already satisfies K(x; x) = 1. This does not mean that normalization of the input features to unit vectors is redundant: Our experience shows that the Gaussian kernel often benefits from it. Normalizing data to unit vectors reduces the dimensionality of the data by one since the data is projected to the unit sphere. Therefore this may not be a good idea for low dimensional data.

**SVM Training Algorithms and Software**

The popularity of SVMs has led to the development of a large number of special purpose solvers for the SVM optimization problem . One of the most common SVM solvers is LIBSVM . The complexity of training of non-linear SVMs with solvers such as LIBSVM has been estimated to be quadratic in the number of training examples, which can be prohibitive for datasets with hundreds of thousands of examples. Researchers have therefore explored ways to achieve faster training times. For linear SVMs very efficient solvers are available which converge in a time which is linear in the number of examples [16, 17, 15]. Approximate solvers that can be trained in linear time without a significant loss of accuracy were also developed .There are two types of software that provide SVM training algorithms. The first type are specialized software whose main objective is to provide an SVM solver. LIBSVM and SVM are two popular examples of this class of software. The other class of software are machine learning libraries that provide a variety of classification methods and other facilities such as methods for feature selection, preprocessing etc. The user has a large number of choices, and the following an incomplete list of environments that provide an SVM classifier: Orange , The Spider, Elefant , Plearn , Weka , Lush, Shogun . The SVM implementation in several of these are wrappers for the LIBSVM library.

**5.2 PROCEDURE**

The input data classify into six output classes. These classes are

* Health
* Sanitation
* Infrastructure
* Employee
* Security
* Administration(Finance)

We further divide the security class into theft security and passengers’ security, infrastructure class into infrastructure of train and infrastructure of platform and sanitation class into unhygienic train, unavailability of water and bad toilets.

**Health**

 The provision of preventive and promotive Health Services are essential to control communicable and non-communicable diseases and to improve the health of the Railway population, so as to enable them to lead a better quality of life. Some of the preventive and promotive Health Services provided on the Railways are:

a) Family Welfare Services  
b) MCI I services including Antenatal care. Immunisation of children and Nutritional supplements, etc.  
c) Control of communicable and non -communicable diseases including implementation of National Health Programmes, like control programmes for Malaria. Tuberculosis, Diarrhoea. Cancer. Blindness. AIDS, etc.  
d) Food Hygiene and implementation of PFA  
e) Monitoring of quality of water supply  
f) Industrial Health  
g) Environmental Sanitation

h) School Health Services  
i) Health Education  
j) Health Services in Fairs and Festivals  
k) Control of Epidemics

Dy. Chief Medical Director (Health and Family Welfare) implements these Community Health Activities at the Zonal Headquarters level under guidance of Chief Medical Director. Medical Officer in charge of Health and Family Welfare in the division is responsible for its implementation at the divisional level. The Community Health Services are given in an integrated manner. All Medical Officers and Paramedical staffs have the responsibility in the delivery of comprehensive health services.

**Sanitation**

Sanitation means maintaining a clean environment so that the beneficiaries stay in neat and hygienic environment. The modern scientific term is Environmental Engineering. Railway stations, colonies and all work places are to be maintained in a hygienic and clean manner and adequate care is to be taken at the planning stage itself. Keeping this objective in view, special emphasis is laid on the collection and disposal of refuse, sewage in a scientific manner. Sanitation services are to be provided in all the railway premises including the railway colonies, railway stations, circulating area, railway yards, office, coaches and the track. Cleanliness of these areas is a multidisciplinary approach by various departments of the Railways viz.. Medical, Engineering, Commercial and Mechanical. The Medical Department maintains the sanitation at railway colonies where Health Inspectors are posted. The Commercial Department maintains other colonies, stations and also the Goods Offices, Parcel offices, etc.. The Mechanical Department looks after the sanitation and cleanliness of the coaches. The Civil Engineering Department does so for the yards, track and underground sewerage areas. A Janitor who has staff working under his control looks after the sanitation in big offices. Sanitation of the bulk of the 8000 Railway Stations on Indian Railway are under Commercial Department.

Sanitation class is sub-classified as

1. Unhygienic train- Cleaning of train is not done on time. Hygiene is not maintained in long distance trains due to which passengers face problems related to health.
2. Unavailability of water in toilets- The water tank of train is not filled timely. Because of this there is always lack of water which leads to unhygienic toilets.
3. Bad toilets- Door locking problem exists in toilets. Toilet seats are not well maintained.

**Infrastructure**

Infrastructure is the underlying foundation or basic framework as a system or organization. It is divided on the basis of infrastructure of train and platform infrastructure of platform.

Infrastructure is subdivided into

1. Infrastructure of train- There is unavailability of charging ports. Lights, fans don’t work properly due to which passengers face problems during summer. Berth seats are torn. Windows are broken and sometimes there is no emergency exit.
2. Infrastructure of platform- Drinking water is not available, less no. of available waiting rooms. Seating arrangement is inconvenient on platform.

**Employee**

 Passengers may face problems due to employee’s misbehaviour. Sometimes employees don’t fulfil their duties on time. Also they are not supportive. Ticket checker might be corrupt i.e. they may take money from passengers to allocate seats.

**Security**

The **Railway Protection Force** (**RPF**) is a security force of [India](https://en.wikipedia.org/wiki/India) entrusted with protecting railway passengers, passenger area and railway property of the [Indian Railways](https://en.wikipedia.org/wiki/Indian_Railways) .This is the only central armed police force (CAPF) which has the power to arrest, investigate and prosecute criminals. The force is under the authority of [Ministry of Railways (India)](https://en.wikipedia.org/wiki/Ministry_of_Railways_(India)). The strength of RPF is about 65,000. As on 1.1.2014, sanctioned strength of Group 'A' officers in RPF is 441 including 127 officers recruited through Civil Services Examination of UPSC. RPF is headed by the Director General (DG). Mr S.K. Bhagat was appointed as DG, RPF on 28 April 2016.

Security class is sub-classified as

1. Security issue with theft- The bags and luggage of passengers can be stolen. The belongings like important documents of the passengers may get misplaced or somebody steals them.
2. Security related with Passengers- The passengers may face any abrupt behaviour against them. Security related to women should be there.

**Finance**

The Finance and Accounts functions are integrated with the executive at all levels in the [Railways](https://en.wikipedia.org/wiki/Railways). At the apex level of policy formulation, the Financial Commissioner, Railways, assisted by Additional Member (Finance), Additional Member (Budget), Adviser (Finance) and Adviser (Accounting Reforms) in–charge of Budgeting, Expenditure, Earnings, Accounting and Accounting Development/Reforms, is there to aid and guide the Ministry of Railways ([Railway Board](https://en.wikipedia.org/wiki/Railway_Board)). At the Zonal level, the General Manager is aided by the Financial Adviser and Chief Accounts Officer along with his assistants.

As the requirement of the project the data is fetched dynamically using the twitter.

Fetching data from twitter :

Twitter data can be fetched from twitter in two ways: a) Rest API b) Streaming Api.  
In today’s blog post we shall go through both using Rest API & Stream API.

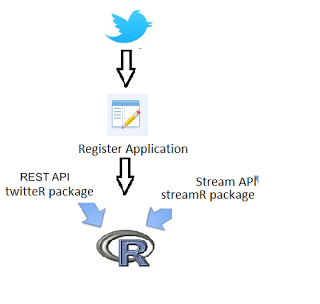
: 

Figure 5.1 - Fetching of twitter data

**twitteR Package:**  
 One of the available package in R for fetching Twitter Data. This package allows us to make REST API calls to twitter using the ConsumerKey & ConsumerSecret code. Code below illustrates how as how to extract the Twitter Data.  
This package offers below functionality:

* Authenticate with Twitter API
* Fetch User timeline
* User Followers
* User Mentions
* Search twitter
* User Information
* User friends information
* Location based Trends
* Convert JSON object to dataframes

**REST API CALLS using R – twitteR package:**

1. Register your application with twitter.
2. After registration, you will be getting ConsumerKey & ConsumerSecret code which needs to be used for calling twitter API.
3. Load TwitteR library in R environment.
4. Call twitter API using OAuthFactory$new() method with ConsumerKey & ConsumerSecret code as input params.
5. The above step will return an authorization link, which needs to be copied & pasted in the internet browser.
6. You will be redirected to Twitter application authentication page where you need to authenticate yourself by providing you twitter credentials.
7. After authenticating , we will be provided with a Authorization code, which needs to be pasted in the R console.
8. Call registerTwitterOAuth().

**StreamR Package:**  
 This package allows users to fetch twitter Data in real time by connecting to Twitter Stream API. Few important functions this package offers are: it allows R users to access Twitter’s search streams,user streams, parse the output into data frames.

filterStream() – filterStream method opens a connection to Twitter’s Streaming API that will return public statuses that match one or more ﬁlter predicates like search keywords. Tweets can be ﬁltered by keywords, users, language, and location. The output can be saved as an object in memory or written to a text ﬁle.

parseTweets() – This function parses tweets downloaded using filterStream, sampleStream or userStream and returns a data frame.

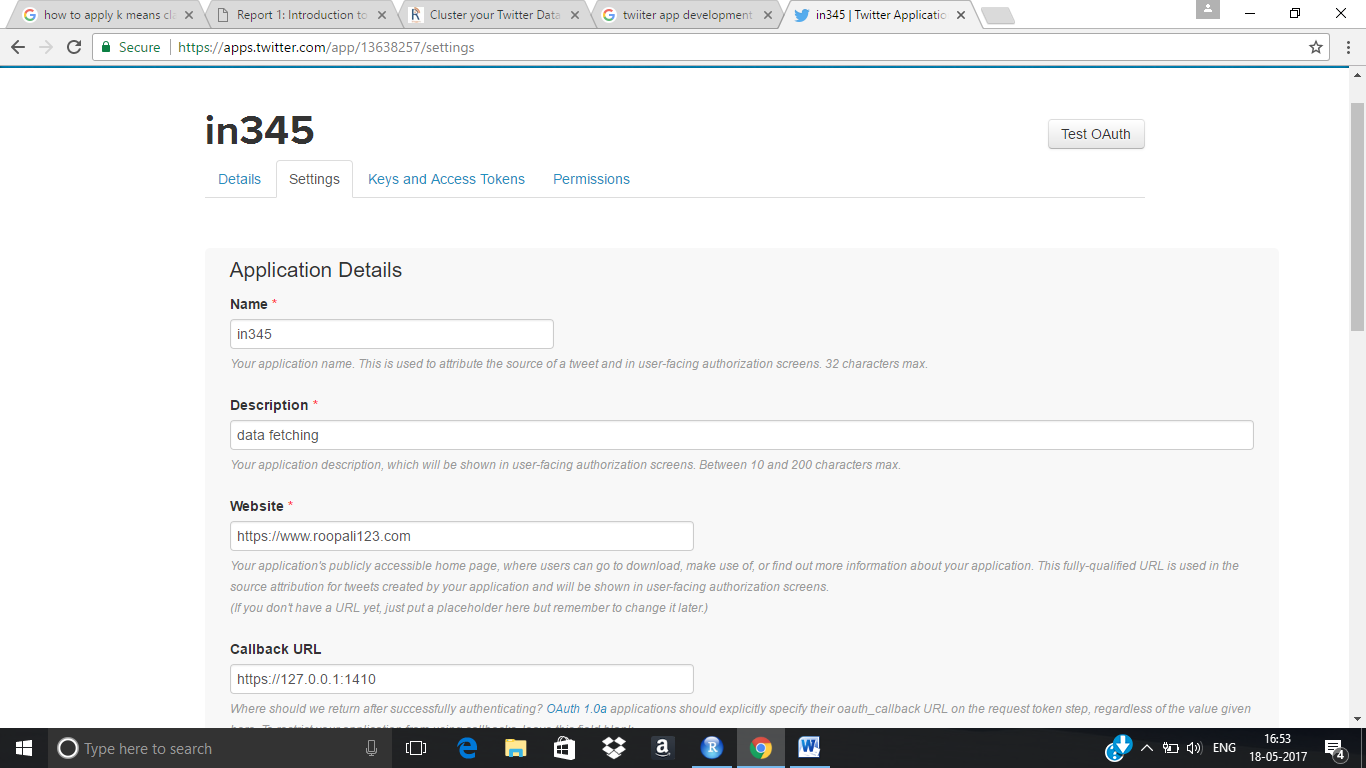


Figure 5.2 - Application Creation

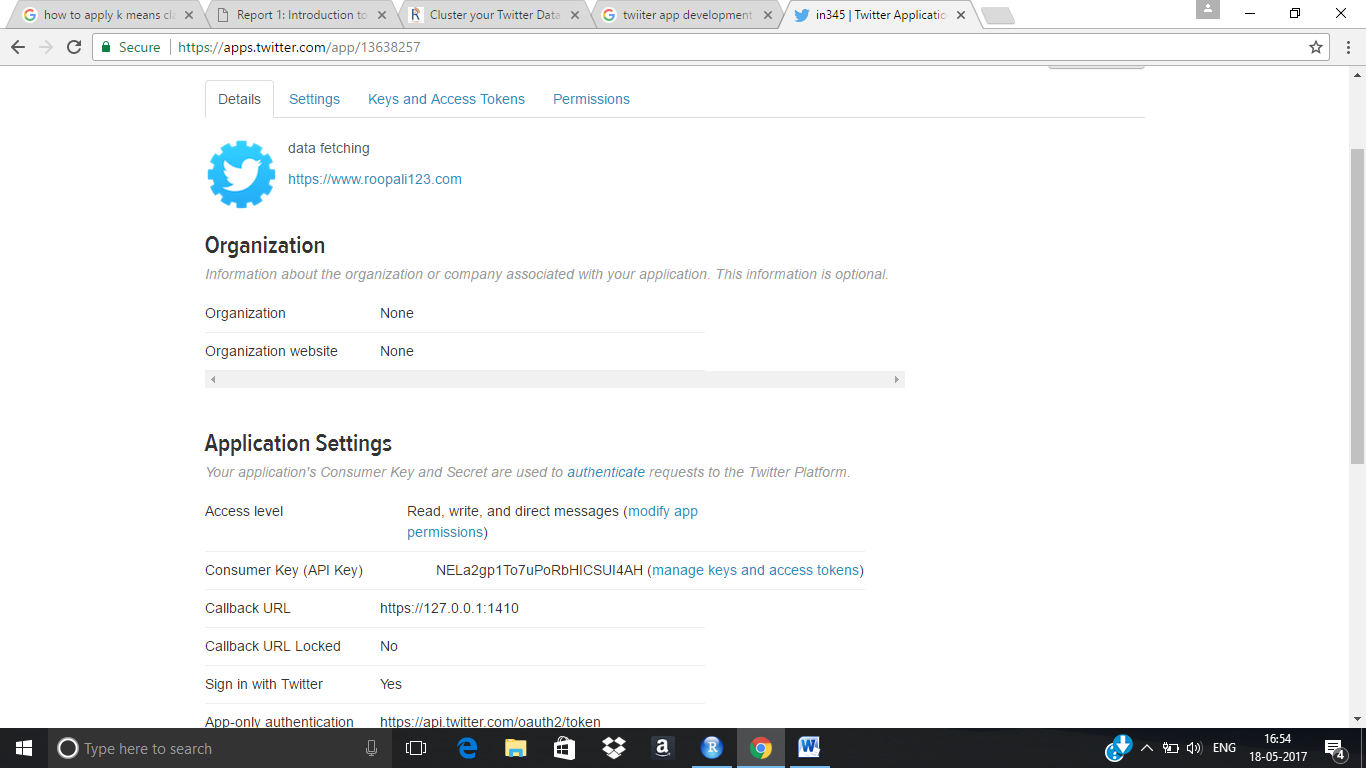


Figure 5.3- Password generation

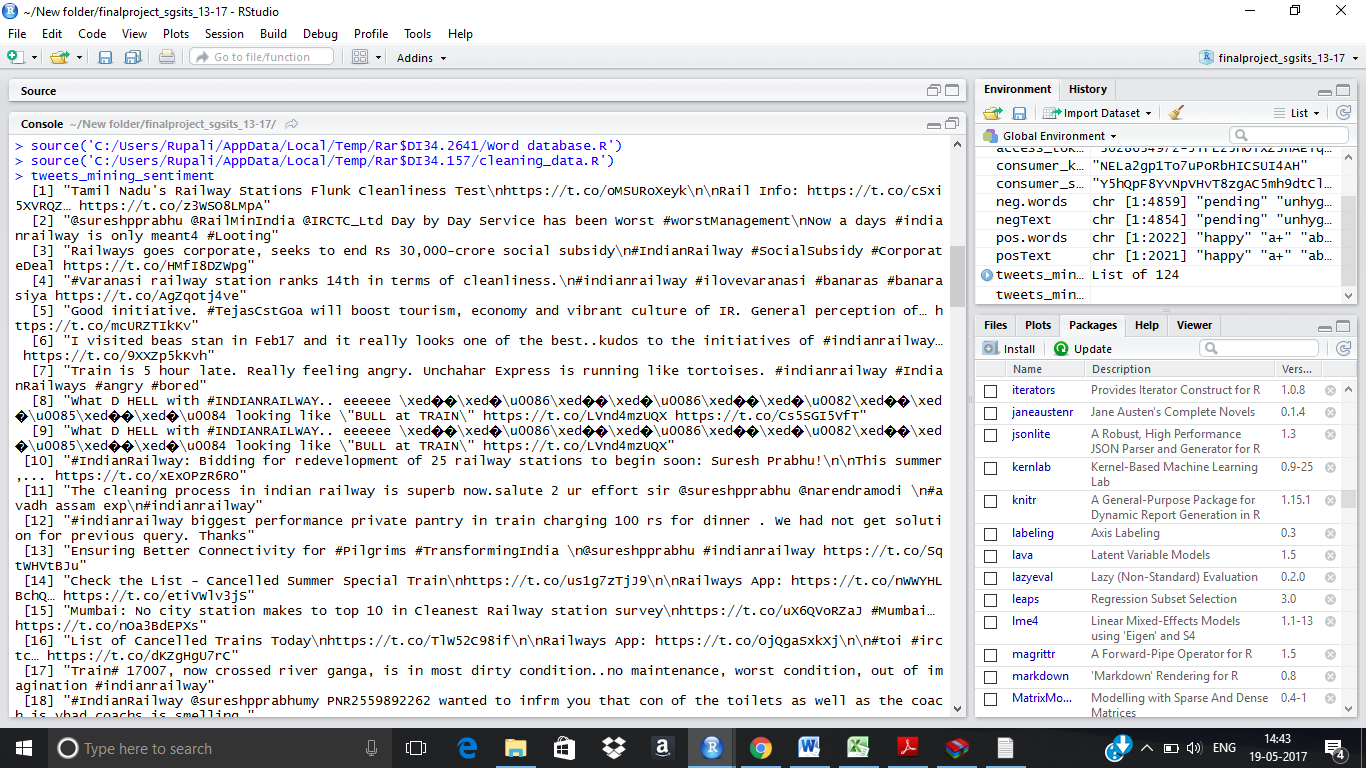


Figure 5.4 – Online Data Fetching

As the tweets contains noises and unrelevant data therefore it is necessary to clean the data as the part of preprocessing. The results of cleaning module is :

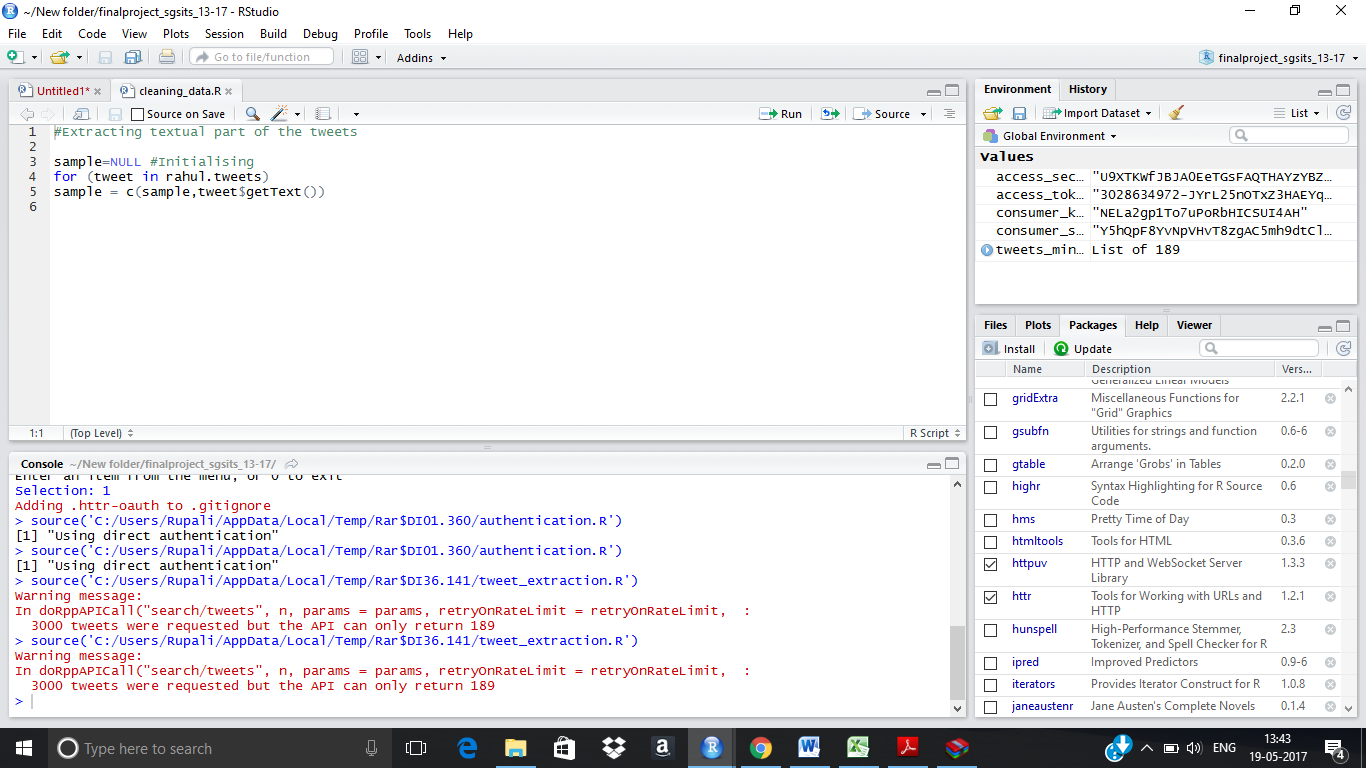


Figure 5.5 - Cleaning Module

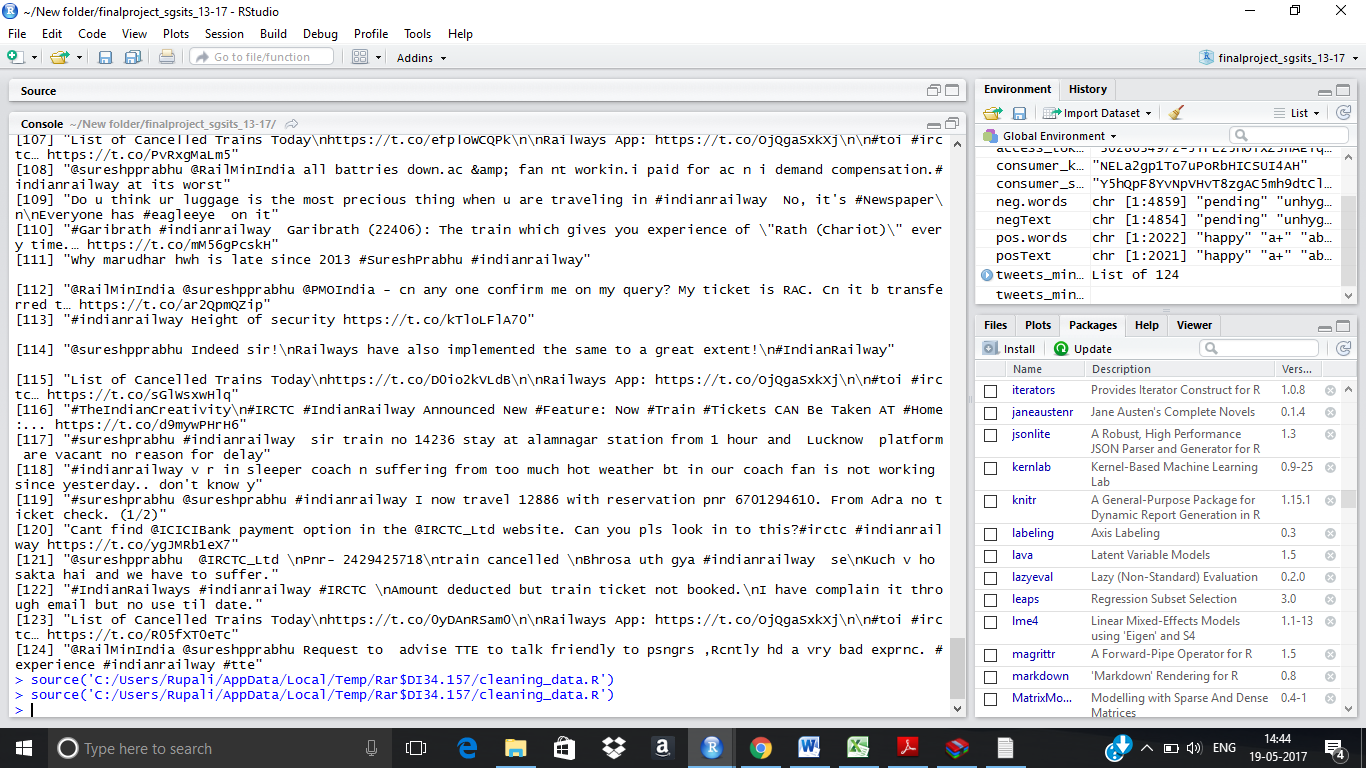


Figure 5.7 - Preprocessed Data

**Training dataset:**

To train the dataset, data from the whole set is partioned into training and test data.

The training set is like:

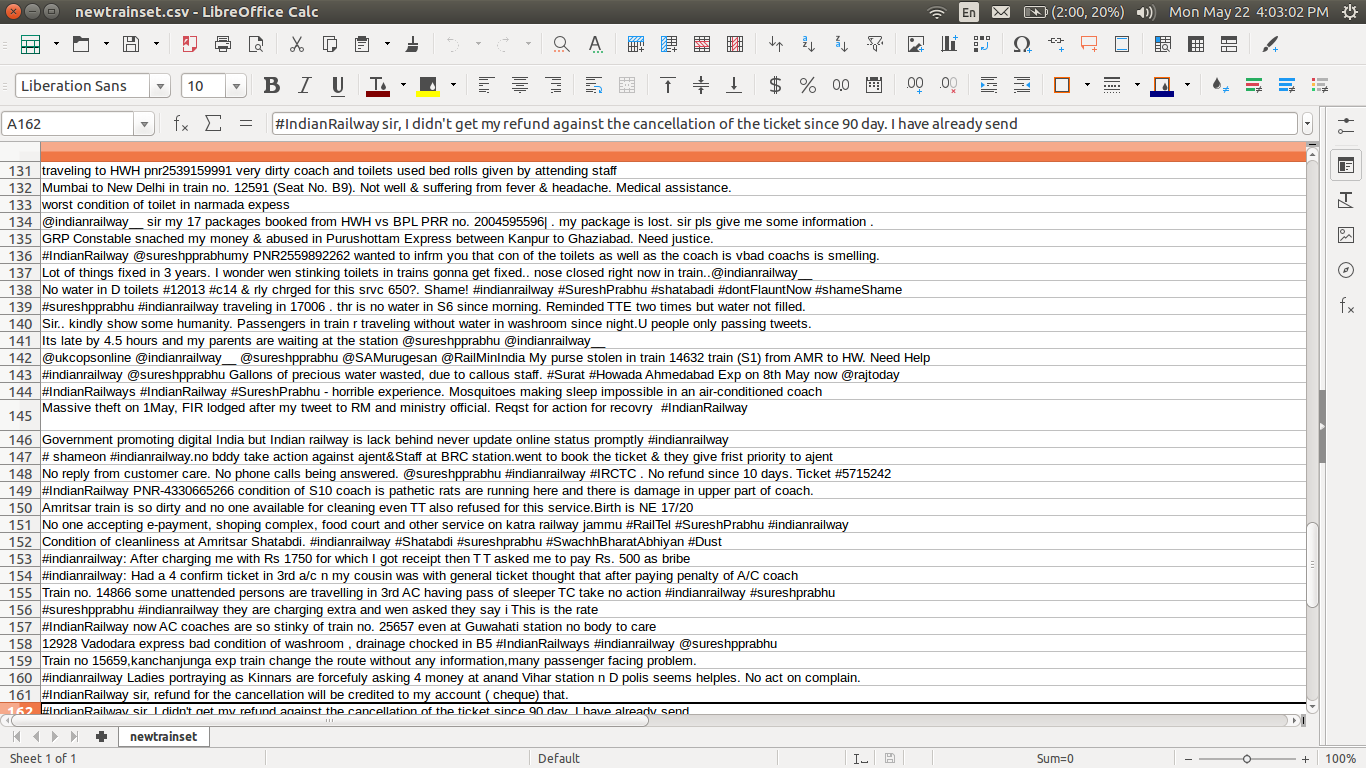


Figure 5.6 - Training Data Set

There are several challenges in our system as the views of user can be dependent on sarcastic manner, as a result we first have to analyse views of users. The first is that a word that is considered to be positive in one situation may be considered negative in another situation.Therefore sentimental analysis have to be performed.

Subset of words which trigger negative polarity of the tweets taken snapshot from the negative file:



Figure 5.8 - Negative Polarity Words

Subset of words which trigger positive polarity of the tweets taken snapshot from the positive file:

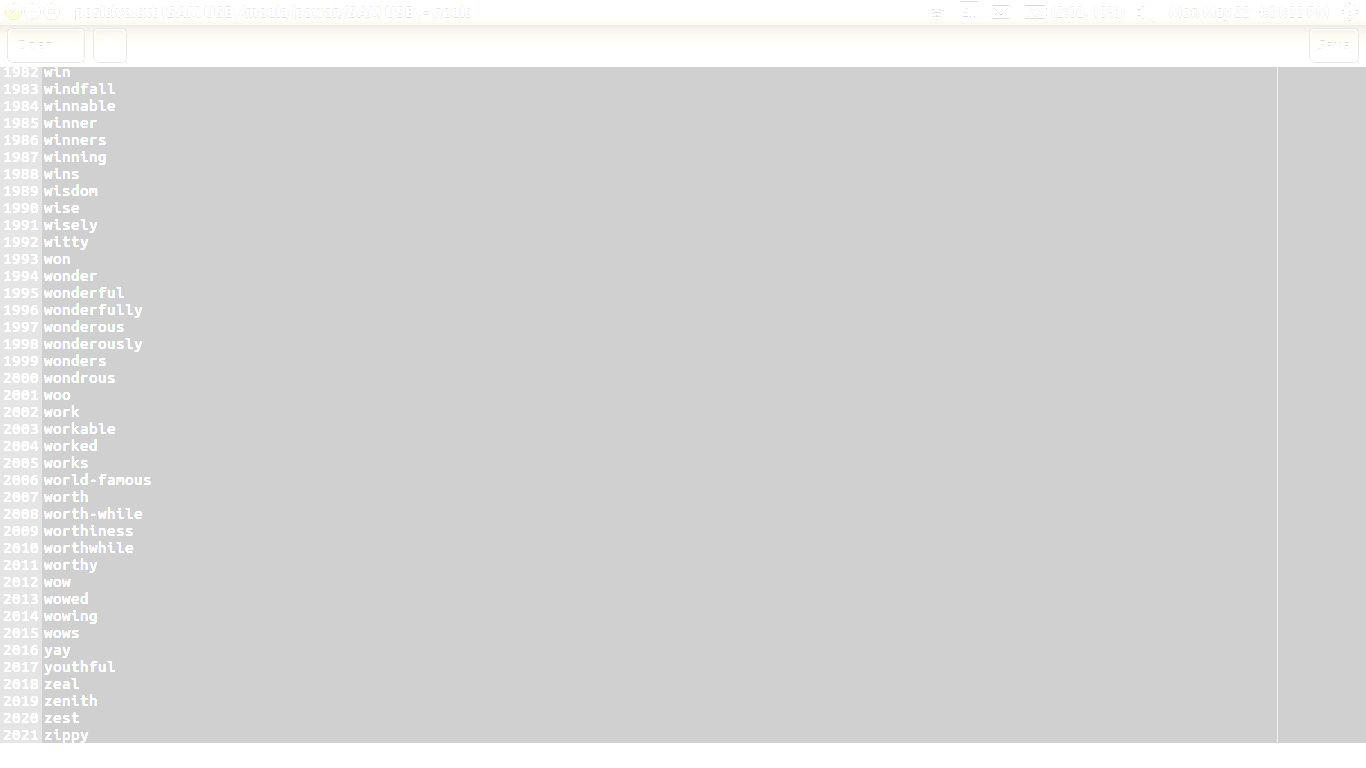


Figure 5.9 - Positive Polarity Words

The result of the sentimental analysis is shown here:

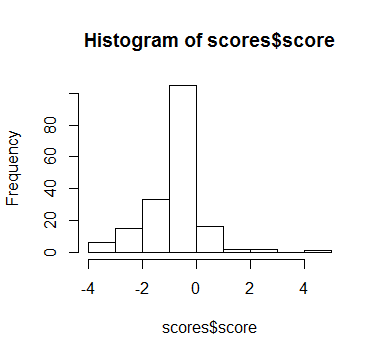


Figure 5.10 - Result of Sentimental Analysis

After the sentimental analysis of data, applying SVM algorithm for the data set and results are as follows:

**SVM OUPTPUT**

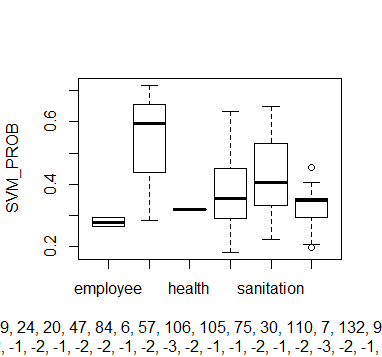


Figure 5.11 - SVM Probability(Level 1)

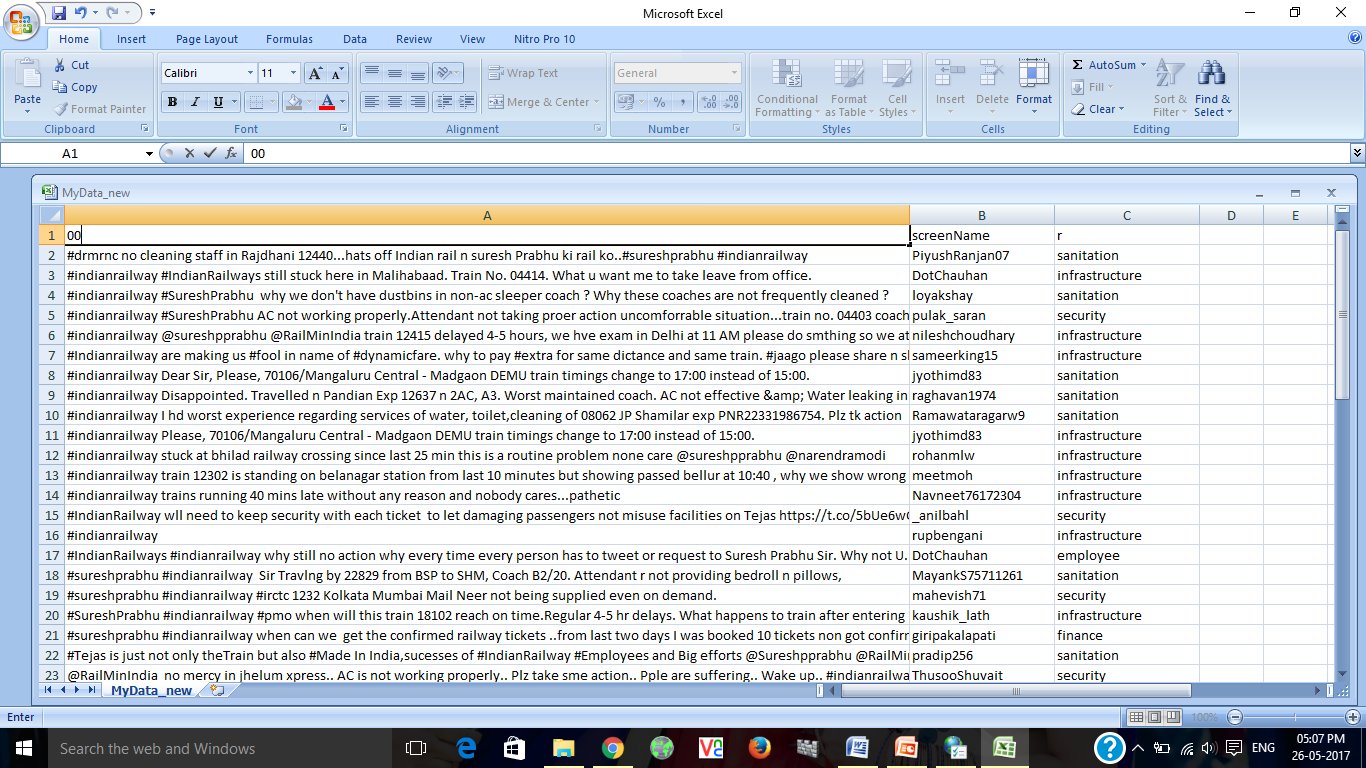


Figure 5.12 Output of SVM

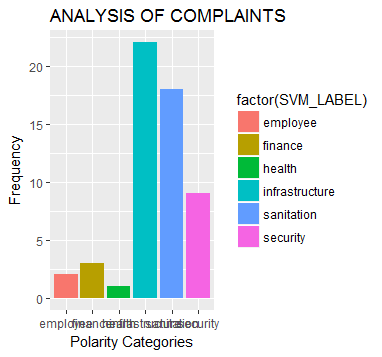


Figure 5.13 - Analysis of Complaints(Level 1)

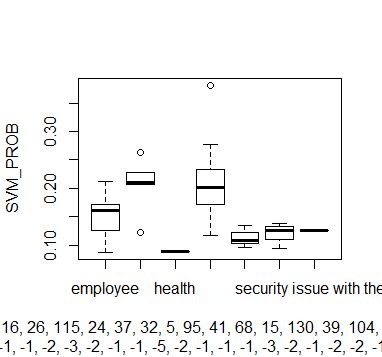


Figure 5.14-SVM Probability (Level 2)

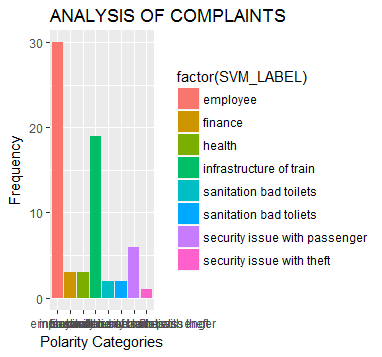


Figure 5.15- Analysis of Complaints(Level 2)

At the current stage, enough amount of data is not gathered. The data that which was used for level 1 classification when used for level 2 classification gets divided into sub parts and thus causing an inaccuracy in the analysis of data and increases the error rate.

This error will be minimized as soon as large amount of real time data is gathered for longer duration of time as this addition of level 2 was just added few days before.

**CHAPTER 6**

**TESTING**

For testing the accuracy of the system, we use many algorithms for analysis of complaint of users of Indian Railway for the same dataset we used Aggromerative Hierarchical clustering (AVC), Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM). After analysing various test cases we found that svm gives best accuracy among all three algorithms.

**TEST CASE 1:**

Hierarchical Clustering :

Aggromerative Hierarchical clustering starts with the point as individual clusters and, at each step, merge the closest pair of clustering by defining suitable proximities.

Result of Hierarchical clustering :

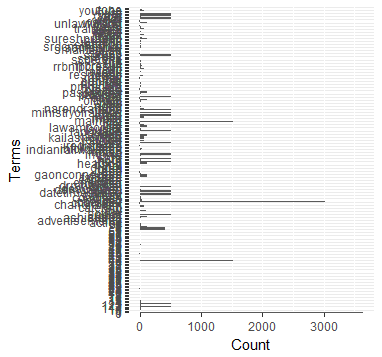


Figure 6.1 - Output of Hierarchical Clustering

**TEST CASE 2:**

**Latent Dirichlet allocation** (**LDA**):

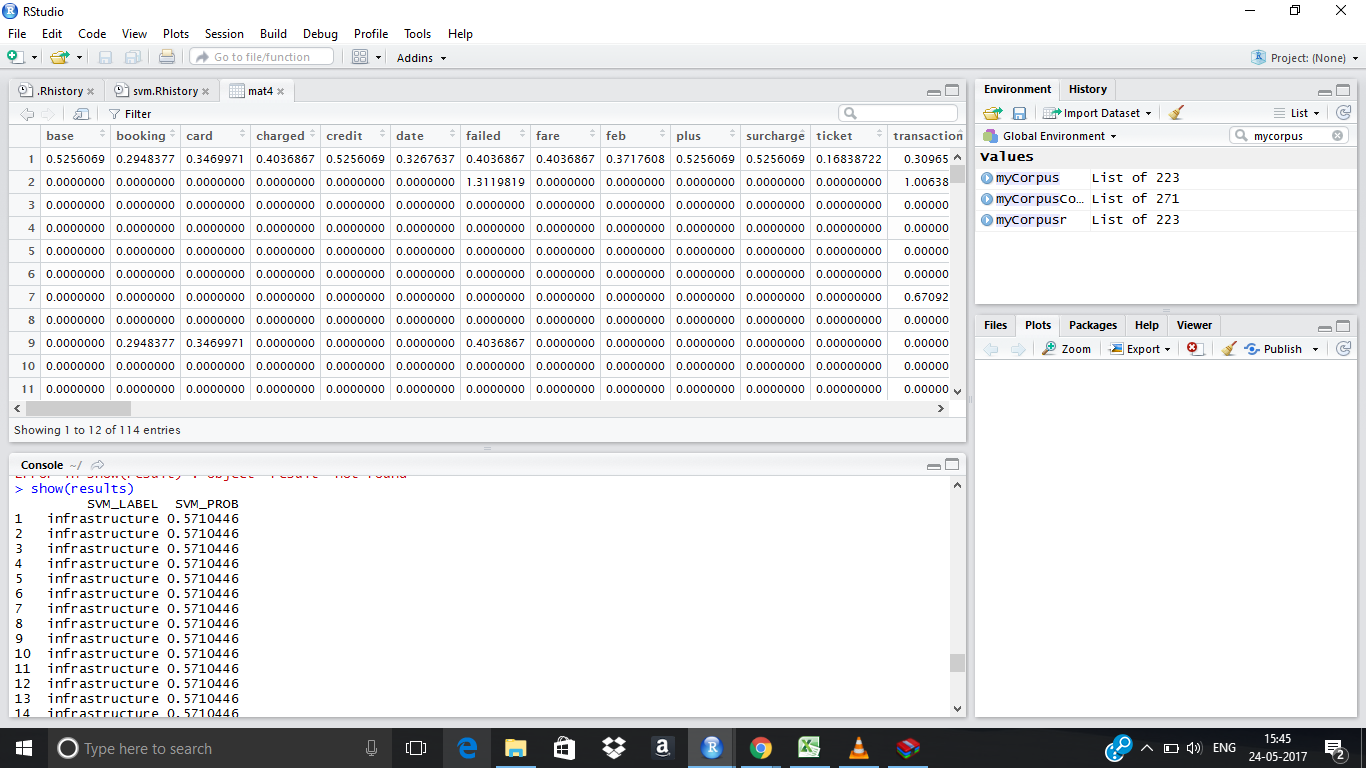
LDA is a [generative statistical model](https://en.wikipedia.org/wiki/Generative_model) that allows sets of observations to be explained by [unobserved](https://en.wikipedia.org/wiki/Latent_variable) groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics

Figure 6.2 - Output of LDA

**CHAPTER 7**

**CONCLUSION**

The analysis of twitter complaints via system includes substantial changes that may entice researchers to understand satisfaction of users towards Indian Railways. The main motive of project was accomplished which was to automate the complaint system of Indian Railways based on the social media, twitter. There exists no system that could automate the system for complaint redressal of Indian railway users . The available system only try to address all the complaints manually.The system operates as a middleware between the users and the authorities of Indian railway. Now the queries and complaints of users are reduced by this system as it analysis the current complaint scenario and predict the frequently occurred complaints. The proposed system eliminates the mentioned problems and takes as input from the twitter page of Indian railway, preprocesses the real time data, analyse the negative complaints, classify them according to training data set and assigns a class for each problem faced by the users and plots a specific graph for the frequent complaint and also forward the immediate complaints to the respective authority. Using the output, the authorities can determine the various lags of the Indian railway system by analysing the user complaints. R Studio was used for the implementation of the project.

**Future Work**

The future work shall begin with the deployment of the software to the Indian railways and enhanced the system accuracy by extracting the more number of tweets from the twitter api . Extracting more number of tweets will lead to develop of better training dataset which will effectively assign classes to the users complaint. In this project, input source was only twitter data we can also combine data of facebook Indian railway page. It can further be improved by taken emojis into consideration which gives exact interpretation of the situation. Images and other languages of the source data can also be a source of data which was not considered in this analysis.

**REFERENCES**

[1] <https://cran.r-project.org/web/packages/available_packages_by_name.html>

[2] <https://www.rstudio.com>

[3] <https://www.r-project.org/about.html>

[4] <https://support.rstudio.com/hc/en-us/.../201057987-Quick-list-of-useful-R-packages>

[5] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge UP, Cambridge, UK, 2000.

[6] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[7] R Reference Card for Data Mining <http://www.rdatamining.com/docs/R-refcard-data-mining.pdf>

[8] Text Mining, in book R and Data Mining: Examples and Case Studies http://www.rdatamining.com/docs/RDataMining.pdf

[9] London, UK. Springer-Verlag. Han, J. (2005).

Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation. Data Mining and Knowledge Discovery, 8:53–87.