Capstone Report CSE-439

On

Twitter Sentiment analysis Using R

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

Capstone Project Report

Submitted by –

Atul Kumar(11505503)

Satish Kumar(11507530)

Sahil Pathania(11505132)

Adarsh Verma(11505616)

Abhishek Kumar(11505240)

Section – KC297

Capstone Group – CSERGC0297

Submitted to – Richa Sharma

Date – 26/10/2018

Declaration

We hereby declare that the capstone work entitled “Twitter Sentiment Analysis using R” is an authentic record of our own work carried out in B.Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara, under the guidance of Richa Sharma, during January to October 2018. All the information furnished in this capstone report is based on our own intensive work and is genuine.

Name – Atul Kumar, Satish Kumar, Sahil Pathania, Abhishek Kumar, Adarsh Verma

Reg No. – 11505503, 11507530, 11505132, 11505240, 11505616

Atul, Sahil, Satish, Adarsh, Abhishek

Date: 26/10/2018

CERTIFICATE

This is to certify that the capstone project entitled “Twitter Sentiment Analysis Using R” submitted by Atul Kumar, Satish Kumar, Abhishek Kumar, Adarsh and Sahil Pathania, in partial fulfillment of the requirements for the award of Degree of Bachelor of Technology in Computer Science and Engineering at Lovely Professional University, Punjab is an authentic work carried out by them under my supervision and guidance. To the best of my knowledge, the matter embodied in the capstone has not been submitted to any other university institute for the award of any Degree.

Richa Sharma

Assistant Professor

School of Computer Science and Engineering,
Lovely Professional University,
Phagwara, Punjab.

Date : 26/10/2018

ACKNOWLEDGEMENT

We take this opportunity to present our votes of thanks to all those guidepost who really acted as lightening pillars to enlighten our way throughout this project that has led to successful and satisfactory completion of this study.

We are really grateful to Richa mam for providing us with an opportunity to undertake this capstone project and providing us with all the facilities. We are highly thankful to mam for his active support, valuable time and advice, whole-hearted guidance, sincere cooperation and pains-taking involvement during the study and in completing the assignment of preparing the said case study within the time stipulated.

Lastly, we are thankful to all those, particularly the various friends , who have been instrumental in creating proper, healthy and conductive environment and including new and fresh innovative ideas for us during the capstone project, without their help, it would have been extremely difficult for us to prepare it in a time bound framework**.**

Table of Contents

1. Social Networking----------07
2. Online Microblogging----------09
   2.1 Advantages of Microblogging over traditional blogging----------09
3. Twitter----------11
4. Sentiment Analysis----------12
5. Problem Statement----------16
6. Gap analysis----------17
7. Objectives----------17
8. Methodology----------18
9. Preprocessing----------18
   9.1 Collection of data----------18
   9.2 Normalization----------19
   9.3 Repeated Words----------19
   9.4 Removal of stop words----------19
10. Performance measure----------20
   10.1 Precision----------20
   10.2 Recall----------20
   10.3 Accuracy----------20
   10.4 Preprocessing using R----------21
11. Functional Requirements----------21
12. Non-functional Requirements----------21
13. Data flow Diagram----------22
   13.1 Importance of DFD----------22
   13.2 Symbols used in DFD----------23
   13.3 DFD----------23
14. Unified Modelling Language----------25
15. Future Scope----------26
16. Conclusion----------26
17. References----------26

1. Social Networking

Social networking is the grouping of individual into specific groups. It could be apolitical or religious group or group of college students, teenagers, all together sharing information of their interests, mostly online. Twitter, Myspace or Facebook are some of social networking sites that are free of charges and easy to access. This interaction is likely to include friendship, families, group relation and romantic ones. Social networking helps people to make new friends and develop some personal relationships and stay in touch with family very easily. Due to vast number of people connects to networking sites, number of relationships gradually increases. Social networking features combined in one website are: user groups, the latest info about music groups, places for videos and photos, blogs, personal profile, and much more. Social networking sites also helps people for maintaining and developing business contacts contact with them. LinkedIn is the best example for this, as it can be suitable place to talk about business and meet with professionals. It's easier and faster to be involving with new business clients.

Social networking requires user's personal details in order to gain full access to the site as sign up. Recent information and news disclosed that some of the social networking websites misuses the personal information of users. Advertisers evade users' privacy. Sex offenders and criminals often visit the sites to find new victims. Some people mostly young ones for sake of revenge and hate post embarrassing information or photos, will have affect the future socially and mentally. These type of crimes called cyber bullying in social networking makes this much faster and easier, unfortunately sometimes even led to death of teens. The developers made the social networking sites for better communication but people rather addictive to those sites. The traditional face to face socializing is becoming obsolete.

2. Online Microblogging

Online microblogging is broadcast medium that exists similar to blogging. Microblogging is different from blogging as its content normally smaller in both total and actual file size. Microblogs allow users to share small chunks of content such as video link, individual images or short messages, which may be the major reason for their popularity. These small messages are sometimes called micro posts. As with traditional blogging, micro bloggers post about topics varying from the simple theme such as "what I'm doing now" to the particular theme like "most watched movie." Microblogs also used for commercial purposes to promote collaboration within websites, products and services and an organization.

Almost all the microblogging platforms offer features like privacy settings, in which users allow to control microblogs by selective ways of publishing entries along with the interface based on web or giving options of their chosen readers. These may include text messaging and instant messaging, E-mail, digital video and digital audio.

Microblogging is slowly moving into the mainstream. For Example, In the United States of America, Presidential candidate Barack Obama microblogged from the campaign trail using Twitter, one of the most popular microblogging services. Traditional organization of media, like The BBC and the New York Times, have begun to send links and headlines in microblog posts.

2.1 Advantages of Microblogging over traditional blogging:-

Why would anyone want to start posting on a microblogging site? If you've been hesitant to jump on a site like Twitter or Tumblr, here are a few reasons to consider trying them

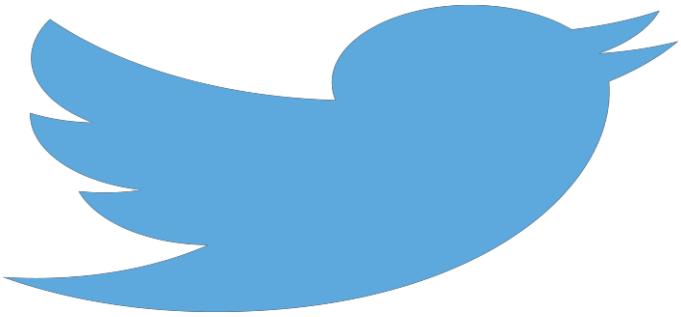
- **Developing content takes less time:** The traditional blogs are quite lengthy so that it takes time to complete our intent. Microblogging gives you the benefit of posting the most recently happened incident to aware their loved ones in shot time and message.

- **Individual parts of the content consumed in less time :** Hence microblogging is such a popular and interactive form of information consumption and social media on mobile devices, because as the gist of the content to the people increases , therefore it is best way where the news comes in short and precise way as compare to long ones that takes time.
- **Increases chances of frequent posts:** Microblogging involves the more frequent posts and shorter ones whereas traditional blogging involves exactly opposite less frequent post and longer. Since you're saving so much time by focusing on just posting short pieces, you `can afford to post more frequently.
- **Share time sensitive or urgent information in a n easier way:** Huge number of the microblogging platforms have been made to be fast and easy to use. With a Vine video, Tumblr post, Instagram photo or simple tweet, you can easily share to everyone on what's happening in your life or any news at this very moment.
- **Communication with followers becomes easy and direct :** In addition to communicate easily with greater short and frequent posts, microblogging platforms can be used to easily encourage and facilitate better interaction through liking, reclogging, tweeting , commenting and more.
- **Convenient using with mobile and tabs:** Microblogging gains too much of attention in present days and the main cause behind this is increasing trends of mobile browsing. It is difficult to consume, interact and write long and lengthy blog post in a tab or smartphone that's why microblogging comes into play and provide small, easy and faster posts.

3. Twitter

Twitter is an online microblogging service that allows users to read and write short sentences of length 140 characters called tweets. Twitter Inc. is located at San Francisco. Users should be register first to post any message, whereas unregistered users can only read them. Users can access Twitter with the website interface, mobile application or SMS. Twitter was created by Noah Glass, Biz Stone, Evan Williams and Jack Dorsey in March 2006 and launched in July 2006. Twitter has 310 Million monthly active user, 1 Billion Unique visits monthly to sites with embedded Tweets, 83% of active users are access through mobile application, consists of 3500 employees around the world, more than 35 offices across the world, 79% accounts are from outside U.S. , supports more than 40 languages and 40% employees of twitter are from technical background. All numbers approximate as of March 31, 2016.

The company experienced rapid initial growth. In year 2007 around 4, 00,000 tweets was posted per quarter. In 2008 this extends to 10 million tweets a quarter. 50 million tweets were posted per day in February 2010. 70000 application were registered by company as March 2010.



According to Twitter 750 tweets posted each second which equals to each day around 65 million tweets were posted as of June 2010. On daily basis around 140 tweets were posted by March 2011. In January 2009, since it gained lot of popularity, Twitter becomes third-highest online microblogging site, given by Compete.com.

The reason we are using microblogging and twitter data are following:

- The scope of microblogging tends to grow bigger and bigger day by day. Easy to use and people can share and give opinions on certain topic, thus it makes essential source.
- Twitter generates vast number of messages that is increasing exponentially. The extracted data can be enormously large.
- Twitter users varies from person to person as user can be politician, film stars, celebrities, sportsman and many leaders across the country including prime minister of India. So it contains all the messages of different caste, religion and sex.
- Twitter users are all over the world so it contains data for different language.

4. Sentiment Analysis

Sentiment Analysis is to determine the opinion of user related to some event or the statement describe the emotion of the user i.e. what he/she feel about it. Users share the things about their ongoing life, discuss current issues and variety of topics. Independent to write in any format without following rules that makes this more popular than older blogging sites. Movies and product reviews easily available now a days or thoughts on religious and political issues, so it becomes essential sources of user sentiment and opinion. Data that we using in our experiment are from twitter, it contains vast number of messages by large number of users created by themselves. Messages can vary from public opinion to personal thought.

These microblogging sites are huge source of information and it is quite easy to say that there is a need of automating the sentiment analysis process as there is too much work involved in processing this information manually. Various approaches are practiced for the automation of this process like machine learning and Natural language processing. Users are increasing day by day as the population and trend of using microblogging sites are increasing, so the data can be used in research purpose of sentiment analysis and opinion mining.

For example, movie makers interested in following questions:-

- What is audience expectation from our movie?(whether the movie is likable or not)
- How the people reacted to our movie?
- Whether the movie is turn to be good or bad?

In the time of election every news channel show the exit polls of every political parties, so every political party willing to know how many are in favor and with the help of microblogging sites people will give the opinions about likes and dislikes of the party. These opinions will help parties to increase their voters.

The data we using in this experiment are movie reviews. We have collected about 17000 movie reviews from twitter. The movie reviews contains reviews of different movies. Reviews can be categorized in three ways:

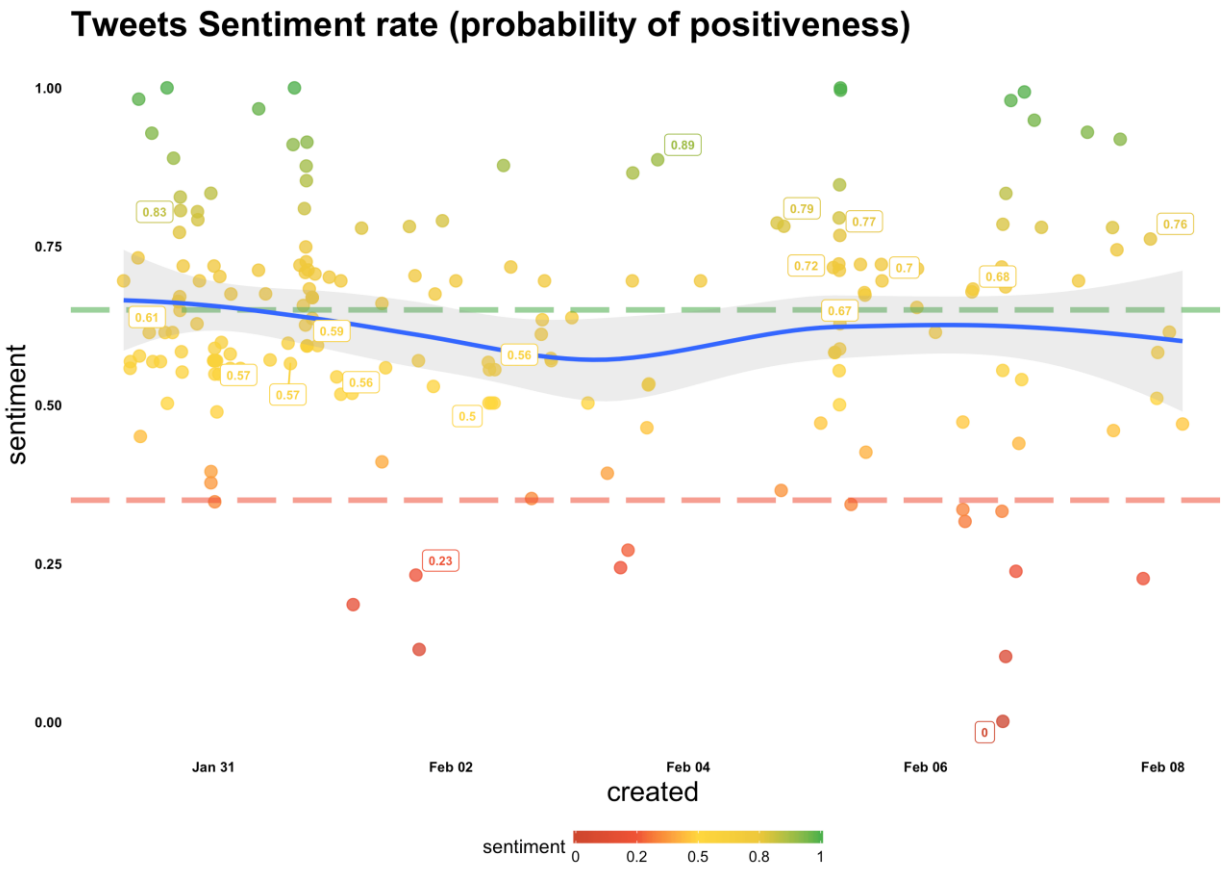
1. Positive reviews: messages in which people liked the movie.
2. Negative reviews: messages in which people not liked the movie.

3. Neutral reviews: messages in which people doesn’t have any emotion or based on mere fact.

We have extracted 5000 each positive, negative and neutral reviews for training set and 2000 reviews will be used in test set. We show comparison between the different machine learning classifiers and find out which will give best results among these.

Table 1.1: Examples of positive, negative and neutral tweets

Sentiment	**Tweet**
Positive	The creators of south park in their own film here, this is a brilliant film with a huge entertainment factor. If you like Naked Gun films and are not young and not too mature or serious on your humor, you'll love this.
Positive	This is the definite Lars von Trier Movie, my favorite, I rank it higher than "Breaking the waves" I simply love the beauty of the picture...the framing is so original; acting is wonderful, A MUST SEE.
Negative	Long, boring, blasphemous. Never have I been so glad to see ending credits roll.
Negative	I rated this a 3. The dubbing was as bad as I have seen. The plot - yuck. I'm not sure which ruined the movie more. Jet Li is definitely a great martial artist, but I'll stick to Jackie Chan movies until somebody tells me Jet's English is up to par.
Neutral	Love to watch commercial movies only in free time but get bored easily.



The green line is the boundary of positive tweets and the red one is the boundary of negative tweets. In addition, tweets are colored with red (negative), yellow (neutral) and green (positive) colors. As you can see, most of the tweets are around the green boundary and it means that they tend to be positive.

5. Problem Statement

Microblogging is type of blogging which consists of limited number of words. Limitation of words determined by respective microblogging sites. It gives right to share his/her thoughts, opinions and sentiments in less number of words. It is one of the revolutionary thing happened in the world of technology. People in these days depends upon microblogging sites such as twitter, Facebook, Tumblr etc. to communicate with both relatives and rest of world. Here sentiments comes into the play which will be shared by anyone in the time they feel and wanted to be shared. Sentiments are nothing but feelings respect to event. Sentiment Analysis is to determine the opinion of user related to some event or the statement describe the emotion of the user i.e. what he/she feel about it.

The research on sentiment analysis has been going for a long time. Sentiment analysis in present days becomes the major issue in field of research and technology. Due to day by day increase in the number of users on the social networking websites, huge amount of data produces in the form of text, audio, video and images. There is need to do sentiment analysis as texts in form of messages or posts to find the whether the sentiment is negative, positive or neutral.

6. Gap Analysis

A lot of research has been done in the area of sentiment analysis. Many researchers used Part-Of-Speech and polarity based feature using supervised learning techniques for classifying. Many automatic classifiers are proposed for classifying the texts in the given expressions but with the restricted domains, but there will be new informal words that are added to the present world which means something in the common social network, so there is need to include all the common referred terms that are used in the social networking world.

7. Objectives

The objectives of the thesis has been discussed in the following points :-

1. To explore, analyze and study the existing sentiment analysis detection techniques in the online microblogging network.
2. To study how the tweets can be generated from the twitter with the help of API.
3. To implement and analyze the results achieved after applying the supervised
4. Learning classifiers to the data set.

8. Methodology

The main aim of the thesis is to compare the results that are implemented with the help of supervised classifier

The methodology followed is:

1. We have collected a corpus of positive, negative and neutral tweets with the help of Twitter4j java API from Twitter. The size of our corpus can be enormously large.
2. We then remove the stop words from the collected corpus to make the content free from commas, full stops etc.
3. We then apply machine learning algorithms to our training set first and then test set and compare the results.

With the help of results we evaluate which machine learning algorithm is best for classification of sentiment Analysis.

9. Preprocessing

a. Collection of data

We collected data from Twitter API named as Twitter4j using net beans. Searched given by using #Hashtag followed by the movie name like #FAN, #Bajarangi Bhaijaan, #The Jungle Book etc. Approx. 17000 tweets have been collected from the various movie tweets.

Reviews can also be searched by #Hash tags followed by respective movie stars, directors, and production house and music record companies. In twitter hash tags becomes the necessary symbol to find about something and it gives user limit of 140 words to express their views and attitude.

b. Normalization

We have found that to get desired results from the classifier we have to make sure that the tweets can be processed properly. As tweets can be in user language, so we have to clean every data which are irrelevant to the data. The following things which can be irrelevant to the data are:-

- URL's: URL's in the message will not make any sense as it simply distracts the result of classifier.
- Username: Removal of username can be necessary for cleaning purposes as it can effect falsely to our results.
- Repeated characters: If the character is repeated more than two time then it can be comprise new word but the meaning is same, so we have to eliminate that word and make the word genuine. For example good can be written as goooooood.

c. Repeated words:

If the message contain word which has been appeared more than two times continuously then it has to be change into two times. For example great great great great movie can be covert to great movie.

d. Removal of stop words

Stop words are the words like "a", "is", "the", "etc" etc; These words has nothing to do with the emotion , so has to be discarded from the message. Now next step is to train the data using supervised classifier.

10. Performance Measure

To calculate the accuracy of classifier we required measure on which accuracy can be obtained. There are two measures on which accuracy can be dependent:

- Precision
- Recall
- Accuracy

Let's take collection of M documents, M_P denotes the number of document which belongs to the true positive class and M_N denotes the number of documents which belongs to the true negative class. TP documents had rightly classified whereas FP documents are wrongly classified, similarly FN documents are wrongly classified and TN documents are rightly classified.

a. **Precision: It is the ratio of documents of rightly classified under positive prediction class to all documents under positive prediction class.**

$$\text{Precision} = \frac{TP}{TP + FP}$$

b. **Recall: It is the ratio of documents of rightly classified under positive prediction class to the documents that are positive in the negative prediction class.**

$$\text{Recall} = \frac{TP}{TP+FN}$$

c. **Accuracy: In order to check which n-gram feature will give better results for these three models, we have to find the accuracy of classifiers. Accuracy for any prediction model can be given as:-**

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

d. Preprocessing using R:

In this step collected data is preprocessed. We have used R language for the preprocessing. Stop words, user references, urls etc are removed from the data. Regular expressions are used to remove url. Collected tweets are then manually labeled and stored in files as test dataset. We have two data sets: positive and negative. We have created two separate files for positive and negative

11. Functional Requirements

Functional requirements are the functions or features that must be included in any system to satisfy the business needs and acceptable to the users. Based on this, the functional requirements that the system must require are as follows:

a. **System should be able to process new tweets stored in database after retrieval.**

b. **System should be able to analyze data and classify each tweet polarity.**

12. Non-functional Requirements

Non-functional requirements is a description of features, characteristics and attributes of the system as well as any constraints that may limit the boundaries of proposed system.

The non-functional requirements are essentially based on the performance, information, economy, control and security efficiency and services.

Based on these, the non-functional requirements are:

a. **User friendly.**

b. **System should provide better accuracy.**

c. **To perform with efficient throughput and response time.**

13. Data flow Diagram:-

Data flow diagram is a graphical representation of data flow in an information system. It is capable of depicting incoming data flow, outgoing data flow and stored data. The DFD does not mention anything about how data flows through the system.

There is a prominent difference between DFD and Flowchart. The flowchart depicts flow of control in program modules. DFDs depict flow of data in the system at various levels. DFD does not contain any control or branch elements.

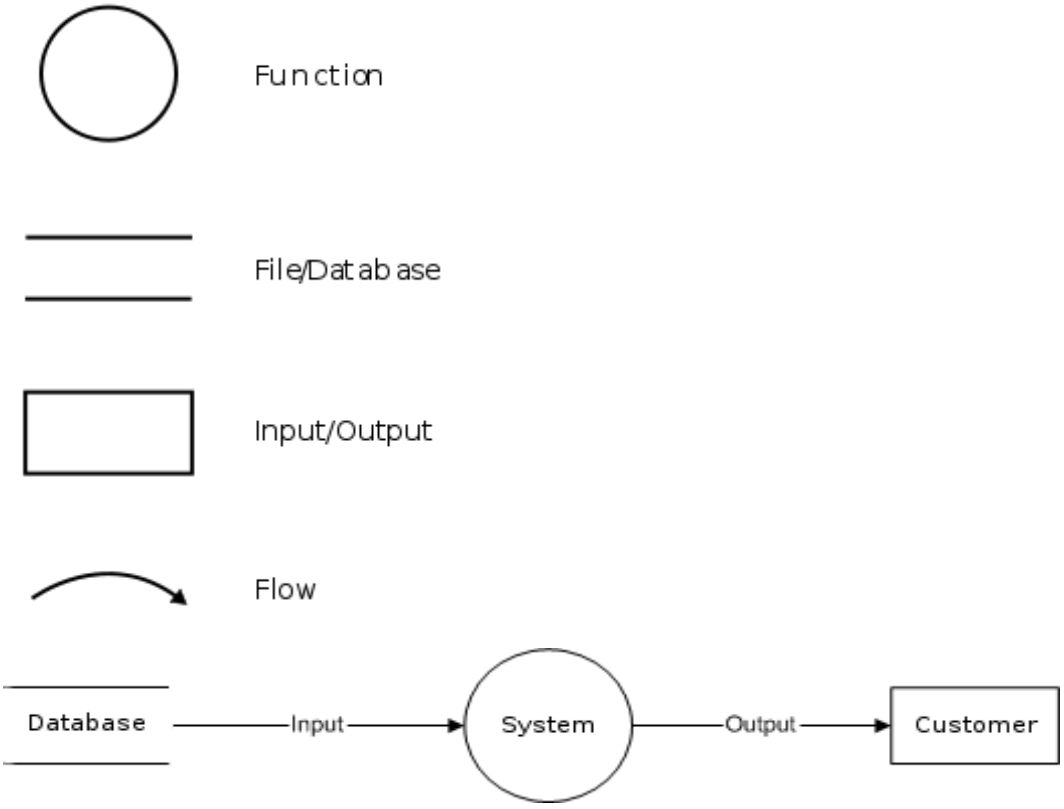
Types of DFD Data Flow Diagrams are either Logical or Physical:-

- Logical DFD - This type of DFD concentrates on the system process and flow of data in the system. For example in a Banking software system, how data is moved between different entities.
- Physical DFD - This type of DFD shows how the data flow is actually implemented in the system. It is more specific and close to the implementation.

13.1 Importance of DFDs

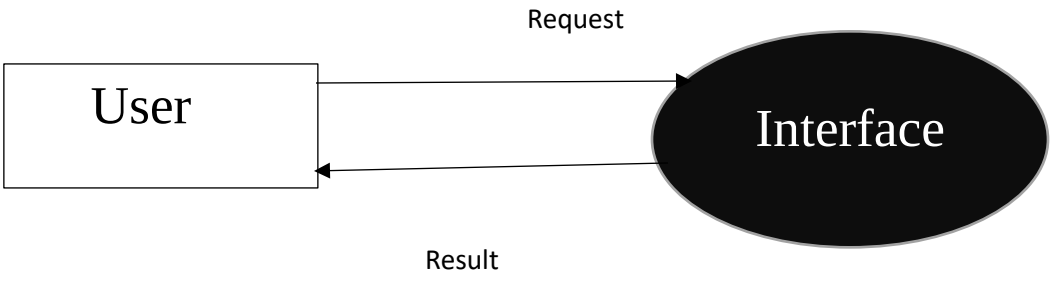
The main reason why the DFD technique is so popular is probably because of the fact that DFD is a very simple formalism – it is simple to understand and use. Starting with a set of high-level functions that a system performs, a DFD model hierarchically represents various sub-functions. In fact, any hierarchical model is simple to understand. Human mind is such that it can easily understand any hierarchical model of a system – because in a hierarchical model, starting with a very simple and abstract model of a system, different details of the system are slowly introduced through different hierarchies. The data flow diagramming technique also follows a very simple set of intuitive concepts and rules. DFD is an elegant modeling technique that turns out to be useful not only to represent the results of structured analysis of a software problem, but also for several other applications such as showing the flow of documents or items in an organization.

13.2 Symbols used in DFD:-

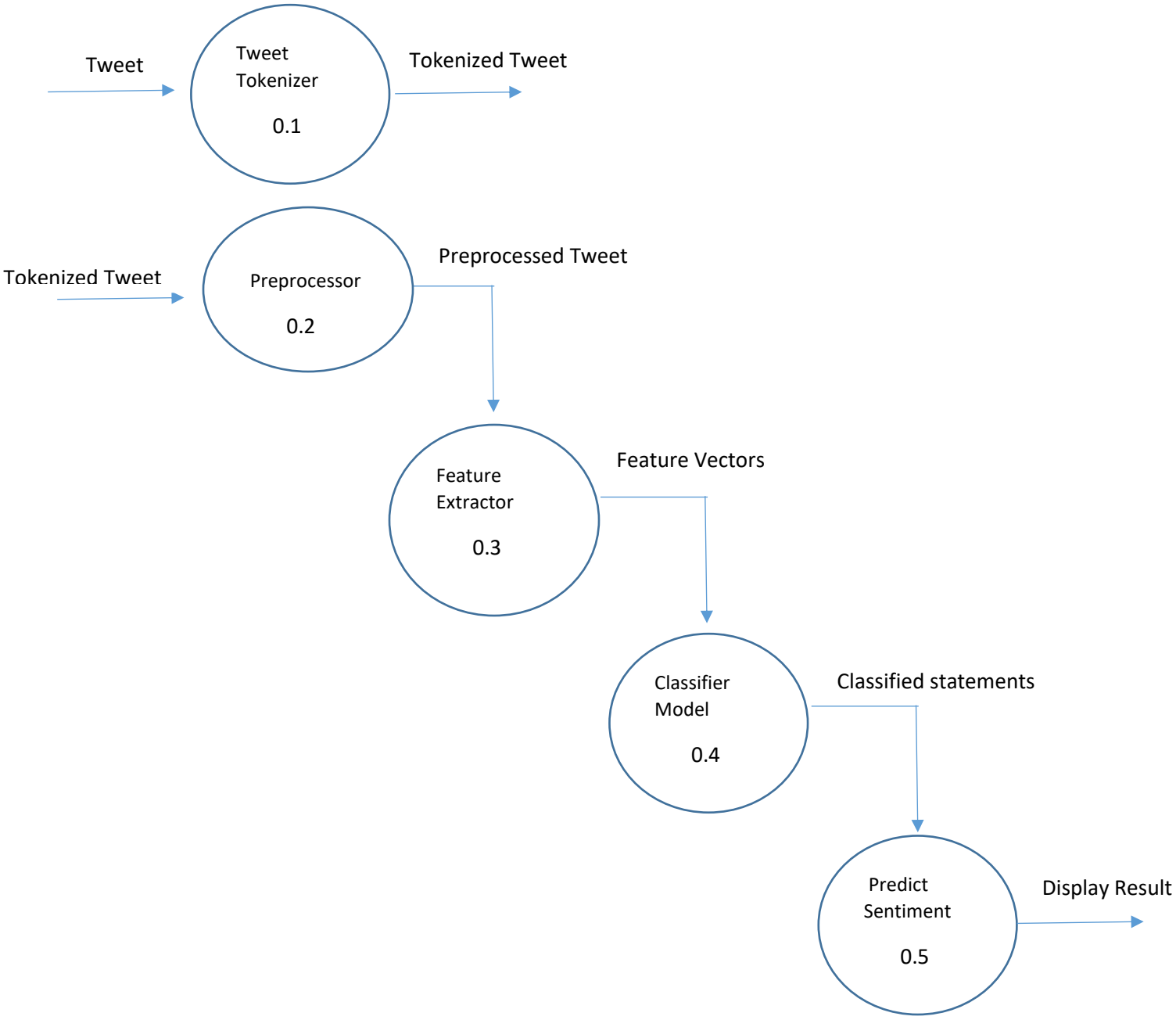


13.3 DFD

Level 0:

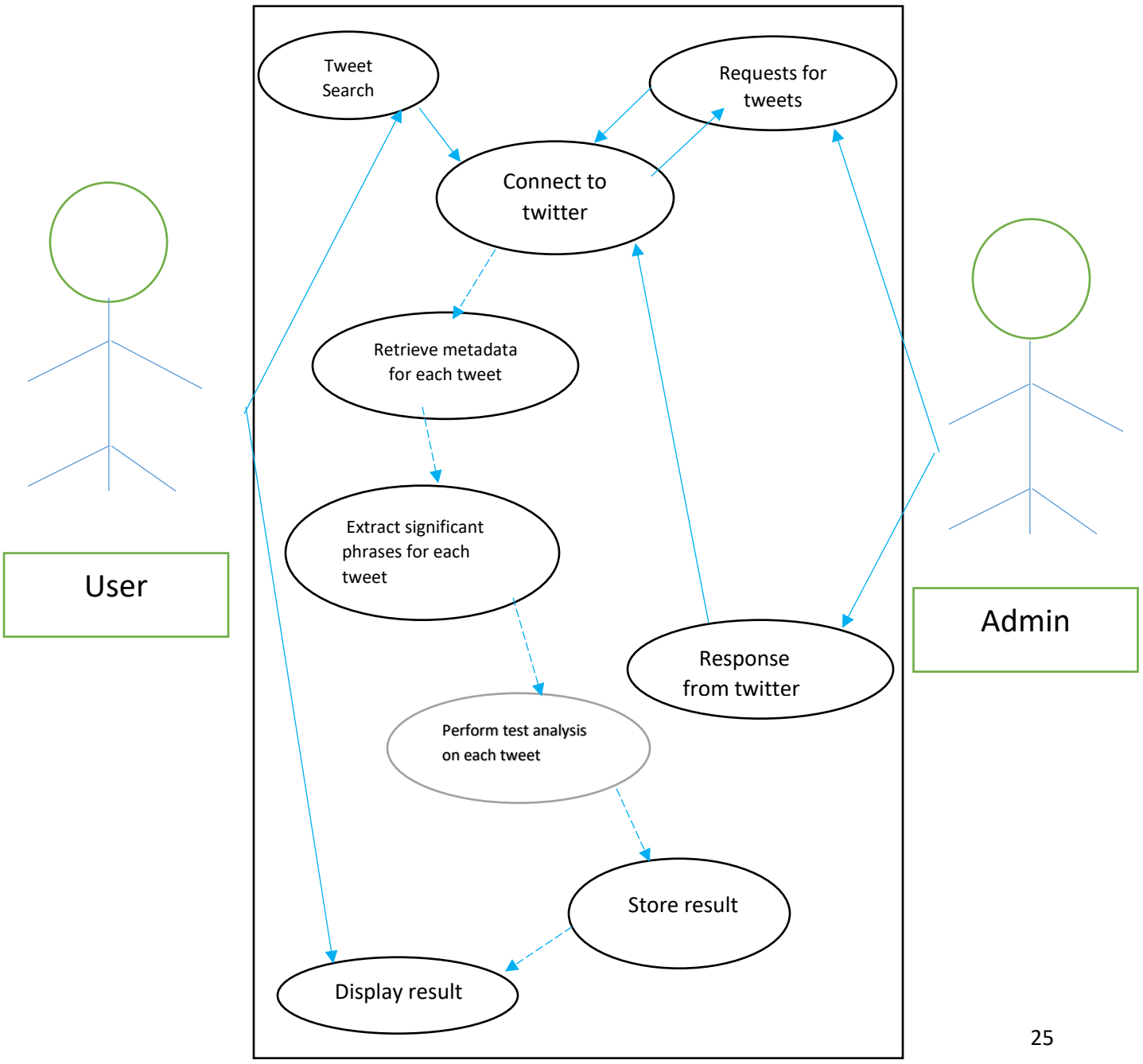


Level 1:


Tweet
Tweet Tokenizer
0.1
Tokenized Tweet
Tokenized Tweet
Preprocessor
0.2
Preprocessed Tweet
Feature Extractor
0.3
Feature Vectors
Classifier Model
0.4
Classified statements
Predict Sentiment
0.5
Display Result

14. Unified Modelling Language:-

UML, as the name implies, is a modeling language. It may be used to visualize, specify, construct, and document the artifacts of a software system. It provides a set of notations (e.g. rectangles, lines, ellipses, etc.) to create a visual model of the system. Like any other language, UML has its own syntax (symbols and sentence formation rules) and semantics (meanings of symbols and sentences). Also, we should clearly understand that UML is not a system design or development methodology, but can be used to document object-oriented and analysis results obtained using some methodology.



15. Future Scope

In future we are planning to make automatic sentiment analyzer for more than one language starting from Hindi language. As nowadays multilingual messages are posted on twitter, so we will able to predict the sentiment for any language.

16. Conclusion

We are going to complete our project using R as a language and the machine learning for output presentation.

In this we took a dataset in order to analyze the tweets. After preprocessing the data we created the feature vector that is used for evaluating Twitter sentiments using Machine Learning techniques. Feature vector includes parameters like hashtags, emotions etc. Knowledge-based approach is used to handle the other words. Slang word frequency count is measured using the backtracking approach wherein the new word is given a sentiment score corresponding to the overall sentiment score of the entire tweet.

We were able to determine the positivity and negativity of each tweet. Based on those tweets we represented them in the diagram like use case diagram, dataflow diagram and through ER diagram. All the diagrams related to outcome are shown in the figure which are mentioned inside the project. A small conclusion is also shown during output presentation based on product or brand entered. Our designed system is user friendly.

17. References

1. https://developer.twitter.com/en/use-cases/analyze
2. https://www.brandwatch.com/blog/understanding-sentiment-analysis
3. https://www.lexalytics.com/technology/sentiment
4. https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17