# Analysis of Users' Interest Based on Twitter Messages

**A DISSERTATION**

*Submitted in partial fulfilment of the requirements for the award of degree of*

**MASTER OF TECHNOLOGY**

in

**COMPUTER SCIENCE & ENGINEERING**


By

**NIMITA MANGAL**




**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**INDIAN INSTITUTE OF TECHNOLOGY**

**ROORKEE – 247 667 (INDIA)**


**MAY – 2016**

# DECLARATION

I declare that the work presented in this dissertation with title, "**Analysis of Users' Interest Based on Twitter Messages**", towards the fulfillment of the requirements for award of the degree of **Master of Technology** in **Computer Science & Engineering**, submitted to the **Department of Computer Science and Engineering, Indian Institute of Technology-Roorkee**, India, is an authentic record of my own work carried out during the period from June 2015 to May 2016 under the guidance of **Dr. Rajdeep Niyogi**, Associate Professor, Department of Computer Science and Engineering, Indian Institute of Technology, Roorkee.

The matter presented in this dissertation has not been submitted by me for the award of any other degree of this or any other institute.

Date:……. Signed:……....

Place: Roorkee (Nimita Mangal)

# CERTIFICATE

This is to certify that the statement made by the candidate in the declaration is correct to the best of my knowledge and belief.

Date:……  Signed:……...

Place: Roorkee

**Dr. Rajdeep Niyogi**

Associate Professor

Department of Computer Science and Engineering

Indian Institute of Technology, Roorkee

# PUBLICATION

1. Nimita Mangal, Sartaj Kanwar, Rajdeep Niyogi, "Prediction of users' interest based on tweets," International Conference on Intelligent Computing and Communication (ICIC2), Kalyani, West Bengal, 18-19, Feb-2016.
2. Sartaj Kanwar, Nimita Mangal, Rajdeep Niyogi, "Event Detection over Twitter Social Media," International Conference on Intelligent Computing and Communication (ICIC2), Kalyani, West Bengal, 18-19, Feb-2016.

# ABSTRACT

Online Social Media now become a part of human life. Twitter is the most famous micro blogging site that is used by major portion of crowd in the world. This site provides us the opportunity to interact with more number of people in less time. Using the twitter data we want to get some useful knowledge that helps humanity. In this thesis, we try to get the interest of users of certain location based on tweets on certain topics (like entertainment, politics, sports, technology, business, etc.). The motivation for this work is to help users to recommending things more accurately. We first analyze the sentiments of tweets and then classify the tweets according to topics. Using any one of the algorithm does not provide us good interest result and hence we combine both the methods. By combining these, we are able to get the topic in which users are positively and negatively interested. We have done this experiment on million of tweets which we have collected for thousands of users and show their interest graph. The results are satisfactory and it validates the proposed approach.

# ACKNOWLEDGEMENT

I would never have been able to complete my dissertation without the guidance of my supervisor, help from friends, and support from my family and loved ones.

I would like to express my deepest gratitude to my supervisor, **Dr. Rajdeep Niyogi**, for his excellent guidance, meaningful insights and moral support. He has been supportive since the day I began working on this dissertation and gave me the freedom I needed to explore this area of research on my own, while pointing me in the right direction in the times of need. His comprehensive knowledge in the area of Social Networking and hard working nature has been a constant source of inspiration.

I am also grateful to the **Dept. of Computer Science, IIT-Roorkee** for providing valuable resources to aid my research.

I would like to thank my friends **Harpreet Kaur, Swarnjeet Kour, Sartaj Kanwar, Sweta Arya, Yashika Jain** who supported me, were always willing to help and give me their best suggestions.

I would also like to thank **Raj Khati Sir** who motivates me throughout the course.

Finally, hearty thanks to **my parents and siblings**, who encouraged me in good times, and motivated me in the bad times, without which this dissertation would not have been possible.

## *Dedication*

*To my parents, for giving me the best education they could*

# TABLE OF CONTENTS

# LIST OF TABLES

Page

# LIST OF FIGURES

CHAPTER 1

# INTRODUCTION

## 1.1 Overview

Now-a-days internet becomes a necessity for human life. This era becomes an era of virtual socialization because of online social sites. Online social media becomes very popular from last 5-10 years. These sites provide us the opportunity to interact with more people and learn more in less time and allow us to be connected with the friends far away from us. Online social sites have now become the fastest means of communication for spreading news to millions of people in a few seconds. Facebook, Twitter, Instagram, FourSquare, Google+, etc. are some of the popular online social sites. Users share their feelings on Twitter by tweets or on Facebook by posting status. Every day huge amount of data is generated by users who are using these social sites. It becomes harder to get some valuable information from this data. Twitter is one of the famous online social blog where many celebrities post tweets for their fans and also post something related to any event occurred. Twitter is a microblogging service. It is so called by this name because it enables users to send and read a short text message which is known as "tweet". It was created in March 2006 and launched in July 2006. There are 316 million monthly active users on twitter and 500 million tweets are posted per day. Since tweet length is restricted to 140 characters so it is a difficult task to predict anything correctly that

based on tweets. These tweets can be used as for analyzing the interest of users in particular locations and get to know about the trends going to that location.

Several works have been done in the field of social networking, which is based on classification of gender, classification of the topic, sentiment analysis of twitter users based on tweets, event detection, community detection, etc. Most of the work on recommendation system is based on network topology. A user's knowledge with social sites service could be remarkably improved if other information like demographic attributes and user's personal interest and the interest of other users was available. Such information allows users to follow a post or user according to his topic of interest and user can join to particular communities of their own interest.

Moreover, a user may be interested to get recommended by things according to her current area of interest. This personal recommendation first requires to knowing the user behavior about which she is discussing. A person gets information about any event through newspaper, television, social sites or with the people around them. Now if a person is interested in that event than she may tweet on twitter about the event positively or negatively according to her viewpoint. To get this negative and positive viewpoint of user sentiment analysis over tweet is necessary. The topic to which a particular tweet is belonged is done by topic categorization and through this we get to know about the topic in which user is interested. By applying both the techniques we can provide better recommendations to users.

In this thesis, we combine both the techniques sentiment analysis and topic categorization and try to find out the current interested topic of users. Sentiment analysis is useful here as it allows us to get the public opinion behind certain topic. It gives us the review for the product that whether people show their interest or not. Topic categorization is useful here because with it, we are able to find out the current discussed topic on social sites.

## 1.2 Motivation

The first motivation for this work is the text length restriction on tweet message to analyze something useful for the user from the short text that user has written on twitter. By analyzing something for the users will provide better recommendations to them. Second motivation is to analyze the tweets of several different places of India and try to find out the most interested topic for that area. This will help the businessman to establish their business in particular area, according to the interest of people over that area.

## 1.3 Problem Statement

The objective of this thesis can be described as below:

"*To design a system that analyzes the interest of twitter users based on their tweets and also analyze the interest of particular location.*"

In order to achieve the desired goal following smaller objectives are set:

1. Collect tweets from the twitter stream for different users and for different locations.
2. Find out tweets contain url or not.
3. Apply sentiment algorithm on tweets and find out sentiment scores.
4. Classification algorithm is applied on tweet and find out the topic to which tweet is related.
5. Finally, evaluate the positive interest under certain topics of particular user or location.
6. Compare sentiment algorithm with modified sentiment algorithm with consideration of emoticon and acronym value.

## 1.4 Organization

The rest of the thesis is organized in the following way:

- Chapter 2 discusses about the related works done in this area. The work done in the area of community detection, event detection, recommendation system, prediction systems, friends' suggestion in a social network by following various approaches proposed by several authors have been discussed.
- Chapter 3 includes the brief description of the algorithms that we are using for sentiment analysis and for classification of the tweet.

- Chapter 4 describes the detailed information of data set and experimental results of the proposed framework. This chapter also consists of comparison result of several algorithms that we are using in our system.
- Chapter 5 presents the concluding part of this thesis and future work which can be carried out for same.

CHAPTER 2

# RELATED WORKS

Due to more use of online social sites by audience makes social network analysis as an interesting topic for research. By applying some data mining techniques, knowledge can be discovered by data of online social sites. This knowledge mostly reflects some trending topics, events, users' suggestion, activities of the user, etc. The problem of social network is considered as a graph theory problem where user is considered as a node of graph and links between users is considered as an edge of graph. Many algorithms for friend recommendation or community detection are based on graph theory.

## 2.1 Use of Twitter Data

Twitter data can be used in many fields for research. Some of them are discussed as below:

### 2.1.1 Sentiment Analysis of twitter data

Different methods are proposed for sentiment analysis, finding sentiments in words, sentences, sentiments in topics. Some of these approaches use machine learning, pattern based and natural language processing. In [1] focuses on four approaches that are already exist and give a different approach for hybrid classifier. First is an NLP based approach in which some existing tools they used such as parsers, N-Grams, POS taggers. The results generated by the tools are grouped in patterns. Each pattern represents some sentimental value either negative or positive. Second is an unsupervised learning in which they form cluster of words and determine the sentiment score for

an expression. Third is a machine learning approach in which they train their system with three different classifier techniques that are support vector machines, ID3, and RIPPER. Fourth is a hybrid classification in which four classifiers are used. First is General Inquirer Based Classifier (GIBC) which depends on only 3672 words which are pre-classified as positive and negative. Second is Rule Based Classifier (RBC) which uses a chunker to parse the sentence and to detect nouns in it and used sentiment value of the word given by pre-classified classifier. Third is a statistic based classifier that calculates closeness between an antecedent represent an expression and a set of sentiment bearing words. They had taken a set of 120 positives and 120 negative words and compute the closeness of words in expression with these sets. Closeness is given by the equation 1 and equation 2 given below:

$$S^{+} = \sum_{n=1}^{120} \text{Closeness}(\text{antecedent}, \text{word}_i^{+}) \quad (1)$$

$$S^{-} = \sum_{n=1}^{120} \text{Closeness}(\text{antecedent}, \text{word}_i^{-}) \quad (2)$$

Fourth is induction rule based classifier (IRBC) which used rule set generated by RBC and SBC and apply ID3 and RIPPER algorithm to it and generate induced rule sets. The experimental procedure that they use for hybrid classification is shown in below figure 2.1.

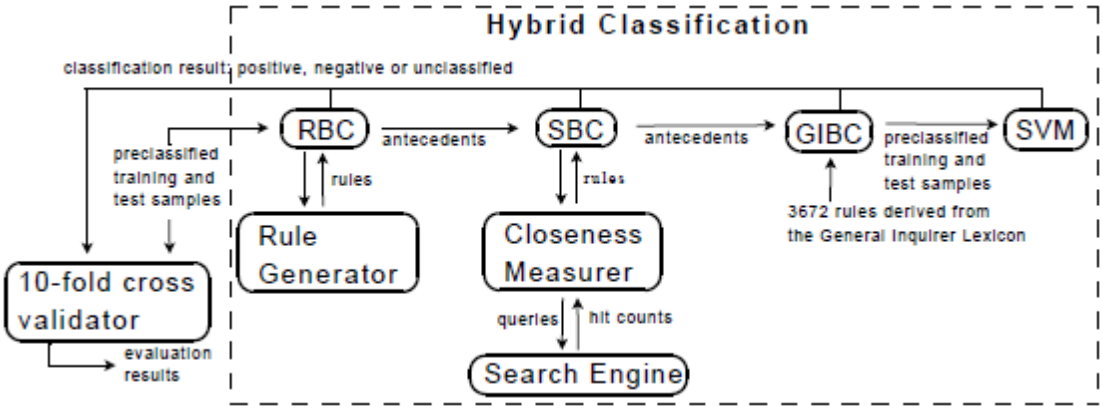


*Figure 2.1*: Experimental procedure [1]

Sentiment analysis of twitter data is studied in [2] and it introduces POS-specific earlier polarity feature and explore the use of tree kernel. Experiments were performed on three models: feature based model uses hundred features only and have the same accuracy as that of unigram model that uses ten thousand features. Kernel tree based model first tokenize the tweet into a tree by separating punctuation mark, exclamatory mark, negation word and emoticon

and prior calculate the polarity of word using word-net dictionary. The unigram model is used as a baseline for the experiments [2].

In [3] two approaches (machine learning and lexical approach) are suggested for sentiment analysis. The machine learning based approach takes text and converts it to a list of words and then takes consecutive pair of words or triplets and calculates some sentiment score based on some code already computed for some set of texts. Using this, new texts are classified into positive, negative or neutral sentiments. In lexical approach, a grammatical structure of language is used and some list of words with sentiment scores and polarities for sentiment score is used. The accuracy of both approaches depends on the training set and the score, which is already provided for most of the words.

In [11] sentiment tree bank approach is suggested for sentiment analysis. The recursive neural network approach computes parent node vectors in bottom-up fashion and use a composition function g and node vector is featuring for that node. In [12] suggested an approach for computing sentiment score of short, informal text and sentence that contain phrases within it.

### 2.1.2 Recommendation System

Many recommender systems provide recommendation using the information based on user profile. [4] suggest a method for user recommendation and the method is based on sentiment volume objectivity. User profiling is done and similarity measure is computed between users (similarity measures based on place, sentiments of tweets). [5] suggest a method for friend's recommendation and uses collaborative filtering and graph structure. In [6] semantic user modeling has been done based on twitter posts. They suggested a formula for user's similarity which is based on topics discussed by the users.

[8] analyze the user intentions that are associated at a community level and show how users with similar intentions connect with each other. [9] address the task of user classification in social media using the machine learning framework. User profile features such as followers, friends, username, user-location are collected to know about a user. Tweets of user are collected for judging the behavior of users and to classify users of same types.

### 2.1.3 Prediction System

In [21] prediction is done in the German federal election held in 2009. They collected tweets that contain name of the parties represented in German election. Sentiment analysis is done using LIWC on the tweets collected. LIWC is widely used in linguistic and psychology. They focus on 12 dimensions to find out that a user belong to which political party. These dimensions include positive emotion, sadness, anger, negative emotion, future orientation, anxiety, past orientation, achievement, certainty, tentativeness, money and work.

[7] suggest a method to predict which political party a twitter user is interested in; based on certain characteristic of parties like activity, influence, structure and interaction, context and sentiment and then user classification have been done based on Bayesian classification.

### 2.1.4 Event Detection

In [15] they have shown how Twitter can be used in important situations. For this they have chosen for high profile events, i.e. to national security and two emergency events. They have shown that messages, send during these events, shows more broadcasting. In [16] they have presented *TwiBiNG* (Twitter Bipartite News Generator). The authors have created this to help online journalism. The main part of this platform is to generate two bipartite clusters of user intensions. After that they have used LCS (Longest Common Subsequence) along with some user information to separate the useful data from irrelevant data. Using this method would not only generate good news, but also contains less spam.

In [17] they have correlated event detection and clustering. Event detection is similar to aggregated trend changes. They have also applied community detection algorithm to find out popular events. In [19] they have used hierarchical clustering of tweets, dynamic denogram cutting and ranking of result cluster is used to obtain the cluster. In [20] the authors have used bursty word extraction for event detection and they also module for location recognition.

In [18] the authors have mentioned that all earlier work has considered temporal context of messages, but location information is also a very important factor of an event. The Geo referencing used in tweets can be used to detect localized events such as public events,

emergency events, etc. User mostly near to event location message more information as compared to others. So these users can serve as human sensors to describe an event.

### 2.1.5 Trending Topic Classification

[10] propose two methods for classification of the Twitter trending topic-one based on textual information and the other based on the network structure. In text based model all the hyperlinks are removed from the tweet and then a tokenizer removes stop words and delimited character. Since there is a limitation of 140 characters in a tweet, people use acronyms for words and so a vocabulary is used that has the full form of these words (eg. BR is used to represent best regard). The network based approach uses a similarity model to find out the trending topic say X. It searches for five topics that are similar to the topic X and finds out the similarity index. In [14] text categorization method is proposed that uses support vector machines and gives proof both theoretically and logically that svm is well suited for text classification.

Most of the above work is related to sentiments, recommender systems, and trending topic. However, these works do not discuss about a user's interest on the topic being discussed by the users. Till now suggestions are given to the user by making network graph of user activities like computing the follower of user and the pages that user like on sites. Our approach is different from others because we combine the sentiments of users with the topic in which user is interested. We compute the interests of a particular user and the users of a certain location by taking their sentiments (positively or negatively inclined) towards certain topics.

## 2.2 Research Gap

The weakness of [3] is, its not well performs whenever there is discussion on topics which is not more focused like discussion about any political party and text contains sarcasm. Their method doesn't distinguish between different word senses because they do not use grammatical parsing to overcome this parsing can be done based on parts of speech tagging. The lexical approach used in this needs to be modified for specific topics.

In [11] uses a sentiment Treebank approach, but their approach does not consider emoticons value and acronym values. Because of restrictions in tweet length many users use

emoticons and acronym to show their sentiments. If any tweet contains such characters than this approach not give proper value of sentiments.

In [10] there might be a case when a topic falls into more than one category, but their approach fails in multiple label categorizations.

CHAPTER 3

# PROPOSED WORK

In this chapter we explain the methodology we used for sentiment analysis and classification algorithm. Brief detailing of algorithms and some implementation details has been discussed in the following section of this chapter.

## 3.1 Overview of Work

The below figure 3.1 shows the basic flow diagram of our method.

i. We have collected the tweets of the user for knowing the interest according to his tweet.

ii. Sentiment analysis has been done on the tweets that are collected to know the inclination of users, whether he is positively indicated his sentiments over a particular topic or not.

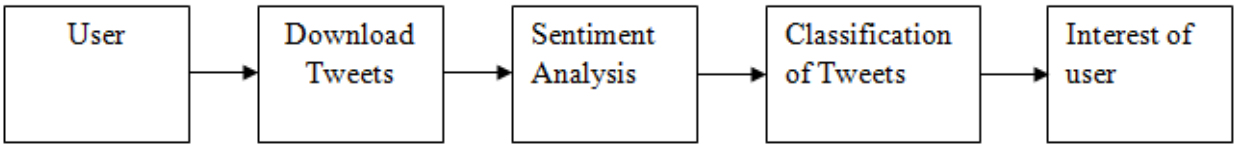


*Figure 3.1*: Overview of our Work

iii. We used supervised algorithm for classification of the tweet to categorize a tweet under a certain label (like sports, politics, entertainment, technology, hospitality, etc.).

iv. Finally interest of the user is given that shows the positive inclination of user towards a certain topic.

For thousand of Indian users, we have collected their tweets, do all other processing and show their interest in the form of a pie chart. In this thesis, we show the result for two of the personalities of India Narendra Modi and Shreya Ghoshal.

We have an example of India in which we have collected tweets of different cities of India and show the current interested topic going on a particular city. We have collected million of tweets and run the algorithm for sentiment analysis and classification over the tweet and get the result in the form of bar graph that shows the current interested topic. By this example we want to show that if any person wants to establish some business in India then which location of India is perfect for him to get maximum profit.

## 3.2 Data Collection

Data is collected for the user for which we want to know the interest and behavior by collecting his tweets, location. We have collected tweets of many cities in India to get major information about that city and to get trend and interest of users at that location. All these tweets are collected using twitter4j API and all the other information about users is also fetched using this API. Twitter 4j is an unofficial Java API for Twitter API. We used streaming API of Twitter 4j for connecting with current tweets related to the user or any geographic location. Streaming API will provide the current tweets only. But the problem with Twitter 4j API is rate limit to access number of tweets per user or per second. Twitter 4j will have rate limit 1 tweet/sec. To access twitter 4j API we need Twitter authentication key which is generated by Twitter developer's option. In other words, we need a Twitter account for generating the authentication key. We have created five accounts on Twitter and generated five authentication keys to resolve the problem of rate limit access and continuously download tweets for our work. We have collected around six lakhs tweets.

## 3.3 Sentiment Analysis of tweet

Sentiment analysis is done by using the Stanford coreNLP sentiment treebank method [11]. This method is appropriate for short text. One drawback of this method, it is not considering emoticons value and acronym value. To solve this problem, first we check for emoticons and

acronyms in the tweet. If it is present we computed the sentiment score accordingly. We generated an emoticon text file in which 107 emoticons are present with their sentiment score positive, negative, or neutral. In table 1 we have shown some of the emoticons and their value. Sentiment value 1 means positive sentiment, 0 means neutral, -1 means negative sentiment.

*Table 1*: Shown the value of some of the emoticons

| Emoticon | Meaning | Sentiment Value |
|---|---|---|
| :) , :-) | Smile | 1 |
| :’(, :,(, :’-( | Cry | -1 |
| :-o, :O | Speechless | 0 |
| :D, :P | Laughing | 1 |
| :(, :-( | Sad | -1 |

We have generated acronym text file only for 50 acronyms which are commonly used in tweets but later it could be expand further. In table 2 we have shown some of the acronyms and their value.

*Table 2*: Shown the value of some of the acronyms

| Acronym | Meaning | Sentiment Value |
|---|---|---|
| ASAP | As soon as possible | 0 |
| BR | Best regards | 1 |
| FCOL | For crying out loud | -1 |
| IDC | I don’t care | -1 |
| LOL | Laugh out loud | 1 |
| K | Ok | 0 |

Sentiment Treebank uses a recursive neural model that has some compositional vector representation for phrases. A tweet acts as a n-gram and this tweet is given to the model. This model will parse tweet and form binary tree and leaf of a tree correspond to word. The value of the parent node is computed in the bottom up fashion using different composition function g.

Figure 3.2 shows the tree structure for very short text "not very good". The leaf nodes of the tree contain words of given text. The parent node is computed by some composition function g. In this case very and good combine and form more positive sentiment but when it combines with not then the sentiment of the text becomes negative.

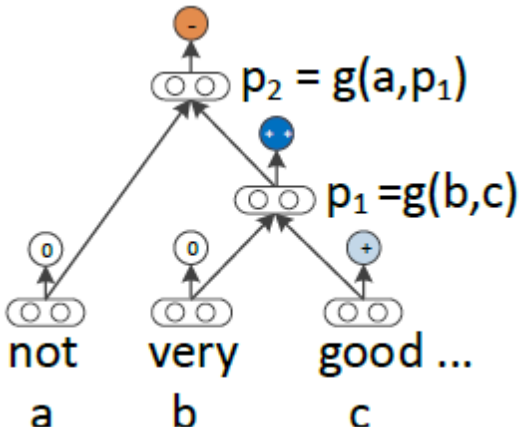


*Figure 3.2*: Recursive Neural network model for sentiment [11]

Figure 3.3 shows the basic flow of sentiment analysis module. First, we replace all the illegal characters like @, extra spaces, etc. from the tweet. We check for emoticon in a tweet, if it is present, then we read the value of emoticon from text file and return the value to the main calling class. Next, we check for acronym in a tweet, if present then compute its sentiment score using the acronym text file. Next, run the sentiment algorithm on the remaining text and compute the final sentiment score. A final sentiment score is given as:

$$Sentiment = a + b + c \tag{3}$$

Equation 3 represents the value of sentiments of different parts which we consider 'a', represents the value of emoticons, 'b' represents the value of acronym and 'c' represents the value of the remaining text. Stanford Core NLP provides some analyzing tools and has techniques to tag the words in a sentence, whether they are names of place, people, etc. or belong to noun, verb, and adjective. These analyzing tools include the parser, sentiment analysis, named entity recognizer, open information extraction tools, etc. First we refine the tweet by removing all hashes, @ and extra spaces to make it more readable plain text. A static init method is called that set the properties to get to know what action is needed for an incoming text. In our case we set four

properties that are tokenize, ssplit, parse and sentiment. Tokenize property breaks the tweet into tokens. The tokenizer saves the offsets of each token from where it starts and ends. Ssplit property split a sequence of tokens into sentences. Parse property generates the parse tree that based on some grammatical structure and language information to distinguish between phrases, subject and predicate in a sentence. Sentiment property uses to compute the sentiment score of a tweet, a binarized tree form for a tweet based on positivity and negativity. After the init method findsentiment method is called that first make a labeled tree for a given tweet and based on tree find the sentiment score in the range of 0-4. Higher the value of the score represents the positive sentiment of a tweet.

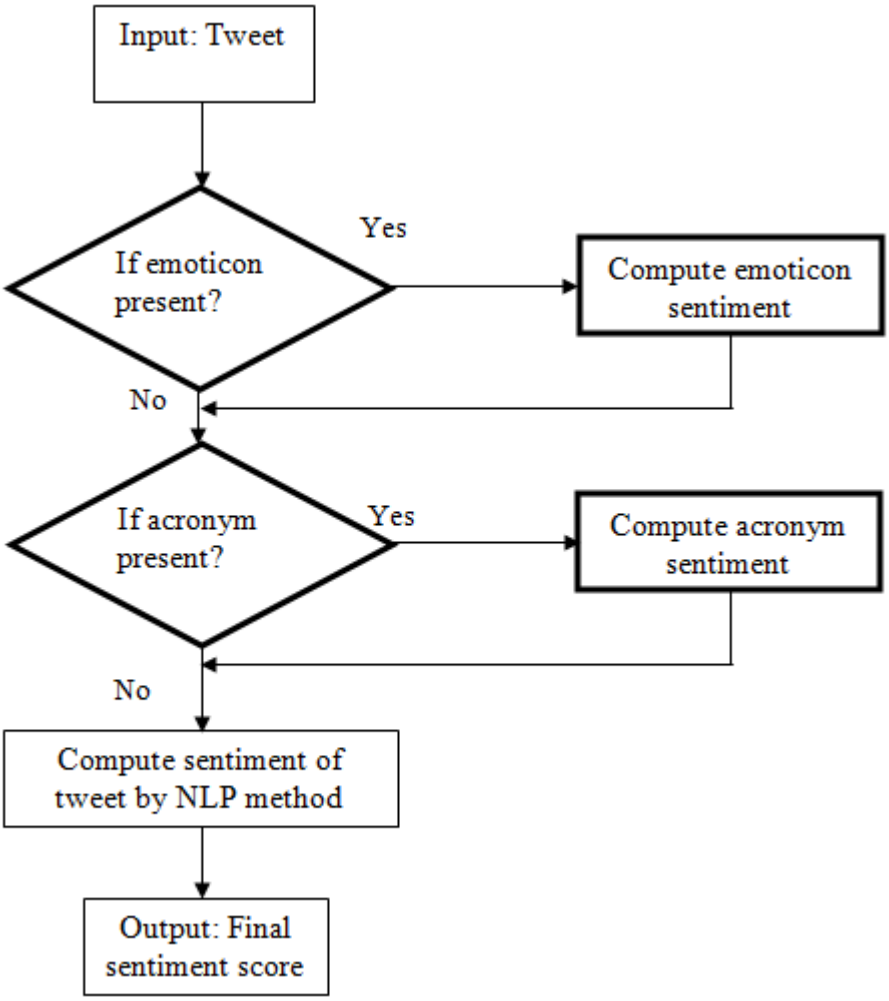


*Figure 3.3*: Flow diagram for sentiment analysis

## 3.4 Classification of tweet

We used supervised algorithm for classification. In the context of the classification the problem can be defined as a set of classes. Here we are considering 10 classes of topic and C is set of classes.

$$C = \{c1,c2,c3,\ldots..,c10\}$$

The classes of topic are predefined and name of the topic is shown in table 3. In addition tweet could be considered as a document and we have to determine the class of each tweet in order to know the interest. We are interested in only those tweets whose sentiment value is either positive or neutral.

*Table 3*: Topic Name

| | |
|---|---|
| Entertainment | Politics |
| Health Medical Pharma | Hospitality Recreation |
| Technology | Society |
| Social_Issue | Business_Finance |
| Sports | Other |

The open NLP package is used for classification [13]. This package provides us a tagger file for tagging of sentences. MaxentTagger is a class used for tagging each word in a tweet with its corresponding form, whether it is an adverb, noun, adjective, etc. There are 36 taggers and each word in a tweet belongs to one of these taggers. After tagging a tweet word tagger pair is formed as shown in figure 3.4. In figure 3.4 first the default properties of trained tagger is loaded. Reading POS tagger model from tagger file and tagging has been done.

```
Loading default properties from trained tagger taggers/left3words-wsj-0-18.tagger
Reading POS tagger model from taggers/left3words-wsj-0-18.tagger ... done [4.0 sec].
Catching/VBG up/RP on/IN movies/NNS ./. ./. Where/WRB else/RB The/DT usual/JJ ./. Mid/NNP air/NN
```

*Figure 3.4*: Shows the word tagger format of a tweet

Each this word tagger pair is compared with ten different categories of topics like entertainment, technology, politics, etc. For comparing word with the topic we are using wordnet similarity

module that implements a variety of semantic similarity and relatedness measures that based on information found in the lexical database WordNet. For using this WordNet similarity we are having WS4J API which having the implementation of this module. For more accurate result we compare these words with the synonyms of topic like, if we want to compare any word with technology then we compare word with technology, network, industry, etc. A method getSimilarity is available which compares this word with these topics and calculates some relatedness scores and gives a similarity score. Figure 3.5 shows the flow diagram for classification of tweet.

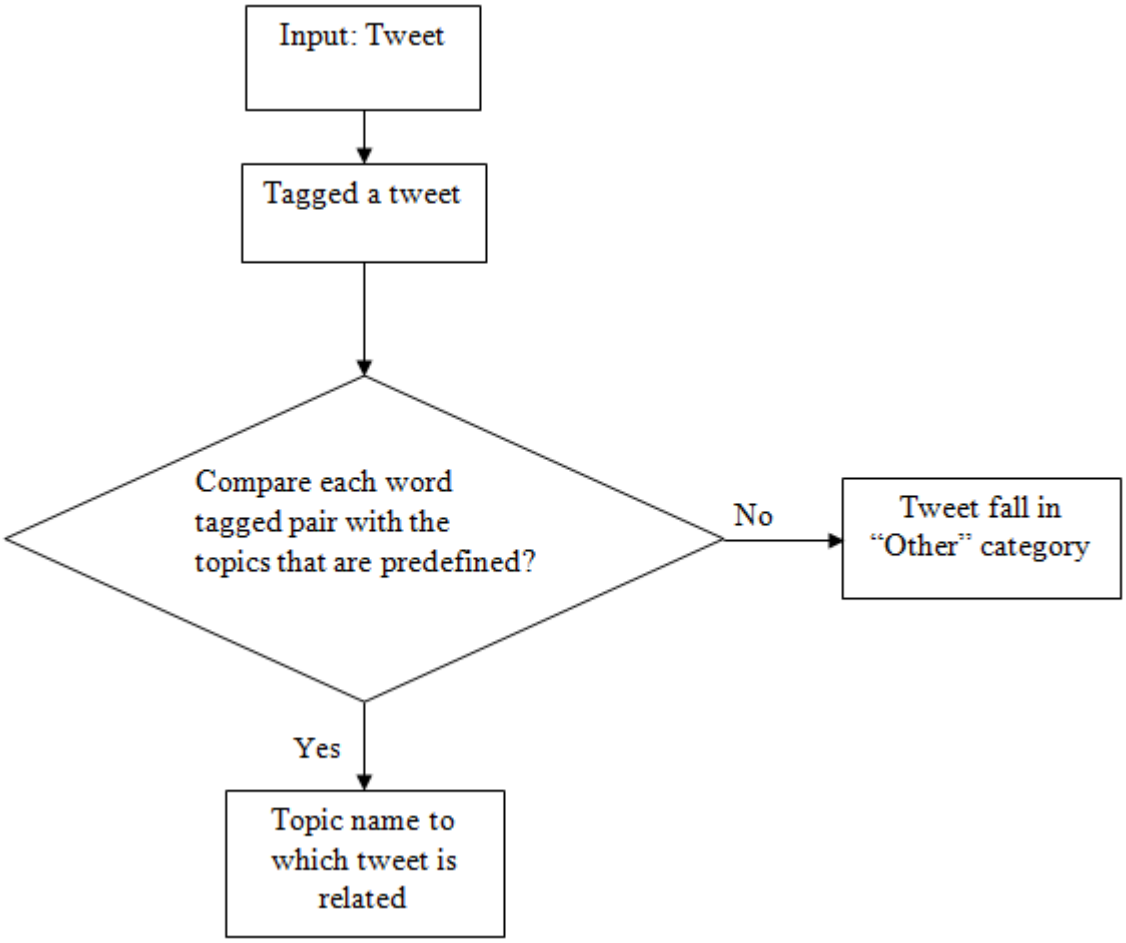


*Figure 3.5*: Flow diagram of classification of tweet

## 3.5 Implementation Details

Below figure 3.6 shows the implemented flow diagram for our system. We have provided an interface to the user in which user provides a screenname (unique name given to each user on twitter) of the user and our backend system calls the download procedure that downloads the tweets of that user and after this sentiment analysis module is called which finds out the sentiment for each tweet and then each tweet is fall under one category positive or negative or neutral. After this classification module is called that runs for the positive and neutral sentiment tweet and gives percentage according to topic to which it belong. It is possible that one tweet belongs to more than one category. The final result for the user about the interested topic is shown in the form of a pie chart.

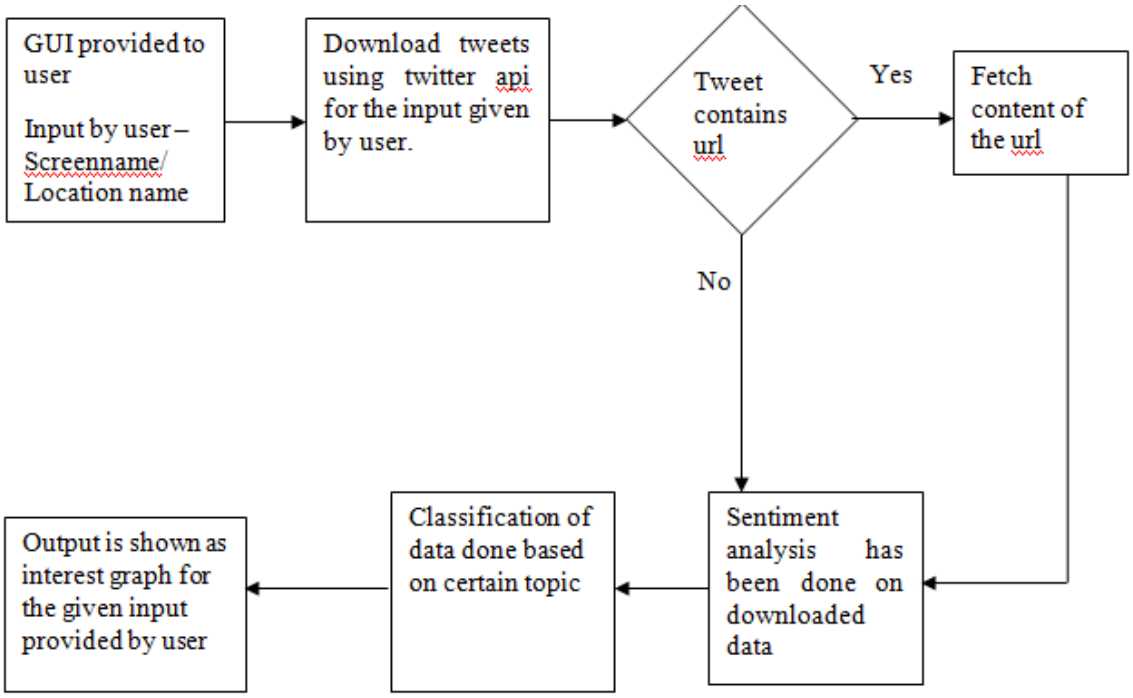


*Figure 3.6*: Implemented flow diagram of our system

If a tweet contains url then first the content of the link is fetched using some code for content fetching of the url and then on obtaining the plain text from url we apply sentiment analysis and classification algorithm.

### 3.5.1 Download Tweet

We first generate the authentication key for using twiiter4j api without this key, it is not possible to download tweets from the twitter. These are the several java classes which are used for tweet download and all these classes are provided by twitter4j api:

i. *ConfigurationBuilder*: Its object is formed which first authenticates the admin user by checking the authentication key which we provided in the code.
ii. *TwitterFactory*: Its object is made by calling the build method of configuration builder class.
iii. *Twitter*: Twitter factory object is used for generating the instance for the Twitter class.
iv. *Query*: Using query class object we query for the topic, here topic refers to the inpt given by the user who is using our interface and this input is either the name of a person or related to some location name. Tweets are downloaded using getTweets method of the Twitter class and stored in our database.

### 3.5.2 Type of tweet

Now downloaded tweets are checked that whether they contain the url or not. This module is using some of the classes provided by Java library and some helping class written by us.

i. *TypeTweet*: This check has been done by a typetweet class in which there is a method named pullLinks is called. We are having the regular expression in the form of string for url type pattern.
ii. *Pattern*: The pattern is inbuilt class provided by Java library which have compile method that compiles our regular expression provided in the form of string.
iii. *Matcher*: Matcher function is called that tries to find out the pattern in the tweet and returns url string if it is present in the tweet otherwise return null string.
iv. *GetUrlContent*: If a url string is returned, then GetUrlContent class is called which first make the object of the URL class provided by Java library and in the constructor of URL class we pass the url string that we have fetched from tweets. Url connection is made and we start getting content from the url and write this content in some text file. Since we get html tags also in our content file so after this we remove all these

tags using jsoup parse method and finally get the plain text on which other computing is done by modules discussed below.

### 3.5.3 Sentiment Analysis

Sentiment analysis has been done on the data which we get after type of tweet module. First we refine the tweet by removing @, #, etc special characters from the tweet then pass this plain text to sentiment analysis class. These are the java classes which we used for sentiment analysis:

i. *SentimentAnalysis*: In this class we are having static init method which is first called. This init method set the properties which are applied on a text and discussed in the paper in the previous section (Section 3.2).

ii. *EmoticonCheck*: Tweet is passed to this class and checking has been done for emoticon. If it is present class will return the value of emoticon sentiments to sentiment analysis class.

iii. *AcronymCheck*: Tweet is passed to this class and checking has been done for acronym. If it is present class will return the value of acronym sentiments to sentiment analysis class.

iv. *NLP*: Remaining sentiment of tweet is calculated in this class. Find sentiment method is called which first calls the process method with tweet as an argument of Stanford coreNLP class and returns annotation object. This annotation object is then used for making a tree for the tweet.

v. *CoreMap*: its class object calls get() method with the annotation object as an argument and this will return a tree object that having a labeled tree for the tweet. Finally getPredictedClass method is called with tree as an argument and this will returns the final sentiment score for the tweet in the form of an integer. This sentiment score value is finally sent to our main calling class method and main class stores this sentiment score with tweet id in our database.

### 3.5.4 Classification

Classification has been done on the data which we get after sentiment analysis. Each tweet is sent to the classification class. The detailed description of following java classes are given below:

i. *Classification*: It is a class which is having some static blocks and intitially these blocks run through which we get the instance of wordnet similarity for Java which is used for finding similarities between words and different topic of classification which we have predefined.

ii. *MaxentTagger*: The Maxent Tagger class object is formed and in its constructor we passed a tagger file (tagger file contains information about the words and its type to which word is belonging like verb, noun, adjective, etc.). tagString method of Maxent tagger class is called for tagging the incoming text or tweet and this function returns a tagged string. This tagged string is split by whitespace and each word is stored in an array of string. Each word and its tagger are passed to getSimilarity method which computes the relatedness of word with the certain topic and if relatedness scores reached to above threshold value that is greater than ten percent in our case then we store this value in an array.

After comparing with all the topics we get the final result in an array which is returned to the main class. These values of topic interest are stored in our database. Finally interest graph is obtained by all these values that we get.

## 3.6 Proposed Method

i. Obtain a choice from a user.
ii. The user enters a screen name of twitter user about which she wants to know.
iii. A module is run, which downloads current tweets of that Twitter user.
iv. Check for type of tweet.
v. Sentiment analysis has been done on the fetched tweet.
    a. Replace all illegal characters (like RT, #, @, etc.) from a tweet and a plain text is processed.
    b. A labeled tree is formed for a tweet.
    c. The score is computed and tweet is classified as positive, negative or neutral.
vi. Classification of the tweet is done as:
    a. Replace all illegal characters (like RT, #, @, etc.) from a tweet and a plain text is processed.
    b. Each word is then classified as a noun, verb or adjective (assign tagger to each word).

c. Each word is compared with the similar kind of word in wordnet dictionary and score for category is decided accordingly.

d. Final score decides to which topic, tweet is belonging (like entertainment, politics, etc.).

vii. Final processing is done and the result is shown as a pie chart which shows the highest interest of the user.

CHAPTER 4

# EXPERIMENTS AND RESULT

In order to justify our claims we have performed several experiments on the real world Twitter data set. In chapter 4 there is a brief description of dataset and result of different user interest is presented. We have also shown the interested topic of various cities of India and shown the comparison of sentiment algorithm we are using.

## 4.1 Dataset Description

We have collected data for three types of problem using Twitter 4j API and experimental results on this data is shown in this chapter.

- First problem deals with the user interest and for this we collected 2,31,750 tweets of 1,150 users and shown their behavior in the form of pie-chart. Tweets for different users are collected for different span of time period for comparing their interest in different time intervals.
- Second problem deals with the different cities of India in which we collected around 2,02,578 tweets of 19 cities of India and shown the interested topic going on that location. Analysis over tweets for different cities is done for the data collected from 25-02-16 to 18-04-16 time interval. Tweets collected by taking the value of latitude and longitude of the city.
- Third problem deals with the comparison of tweets of two countries India and America. Tweets for the users that act as a bot (bot user is a user whose tweet done automatically by

machines and not by person) like news channel (bbcnews, indiatoday, etc) are collected. The reason to choose bot user is to get more news about the country to get to know about the interest of country.

## 4.2 Experimental Results

This section deals with the experimental results for the problem discussed in above section 4.1.

### 4.2.1 User Interest

Using the user interface as shown in figure 4.1 below we have obtained the following result for the inputted user. Here the inputted user is Shreya Ghoshal, a popular singer in India.

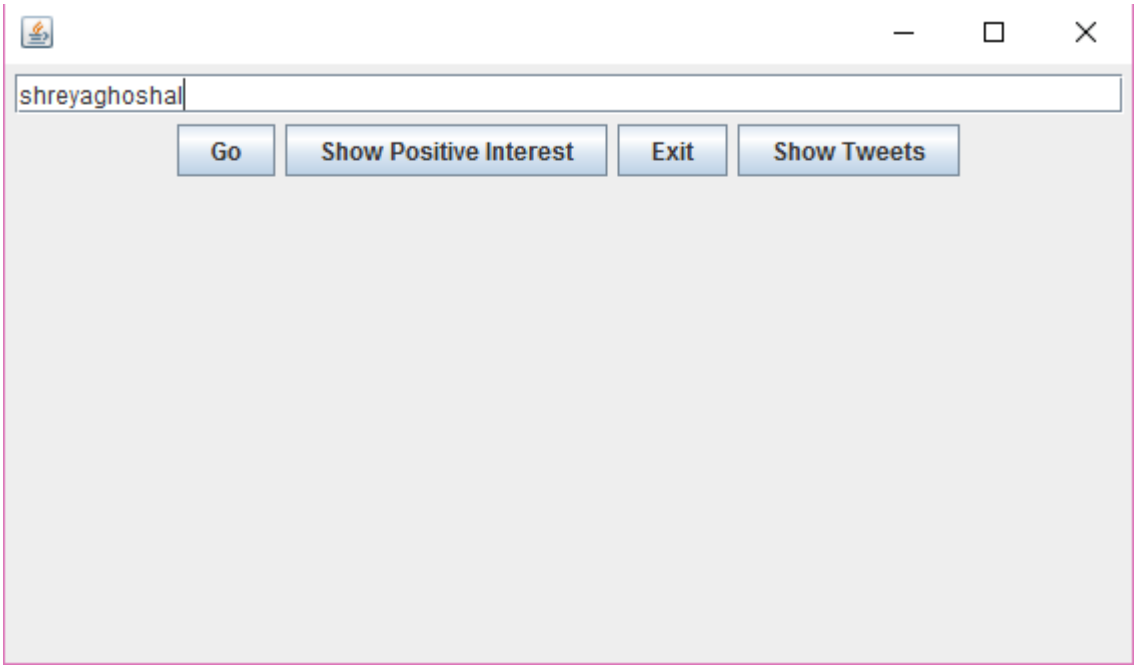


*Figure 4.1*: Interface provided to user

Go button in figure 4.1 is used to download the current tweets of the user of 2-3 days and after this sentiment analysis algorithm and classification tweet algorithm runs on the collected tweet. Finally with show positive interest button we displayed a pie-chart of user interest. Show tweets button display the tweet of user. Figure 4.2 shows the tweets of user.

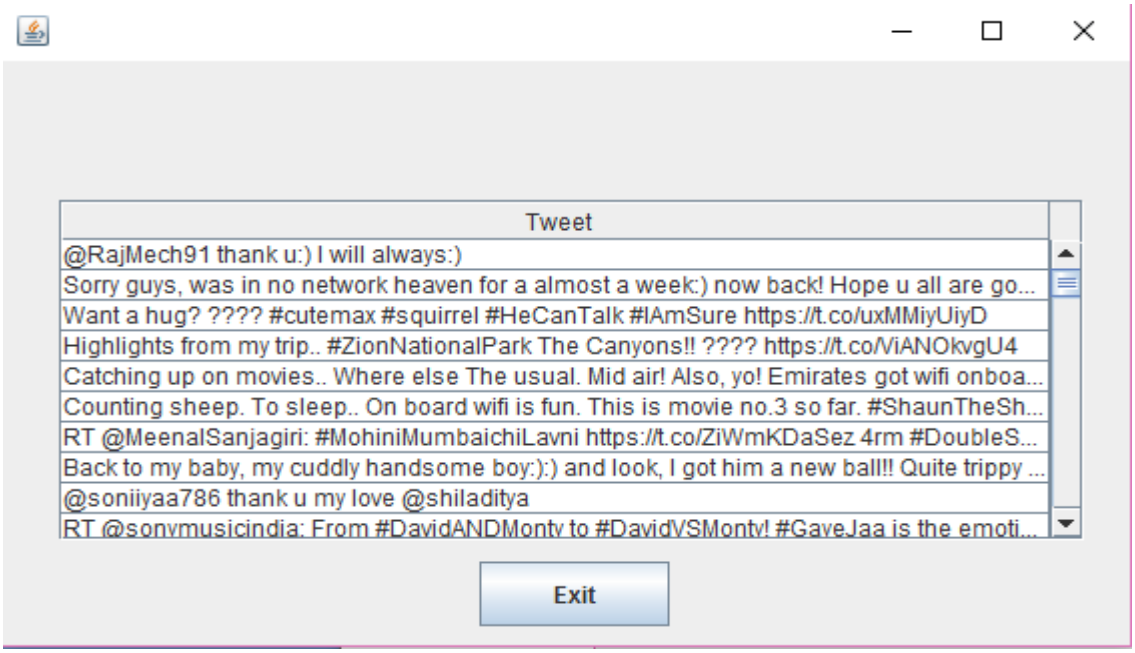


*Figure 4.2*: Show tweets of Shreya Ghoshal

Figure 4.3 shows the interest pie chart for the tweets done by Shreya Goshal, from 1-02-16 to 29-02-16 and this result shows that major topic in which she is interested is entertainment for this period of time.

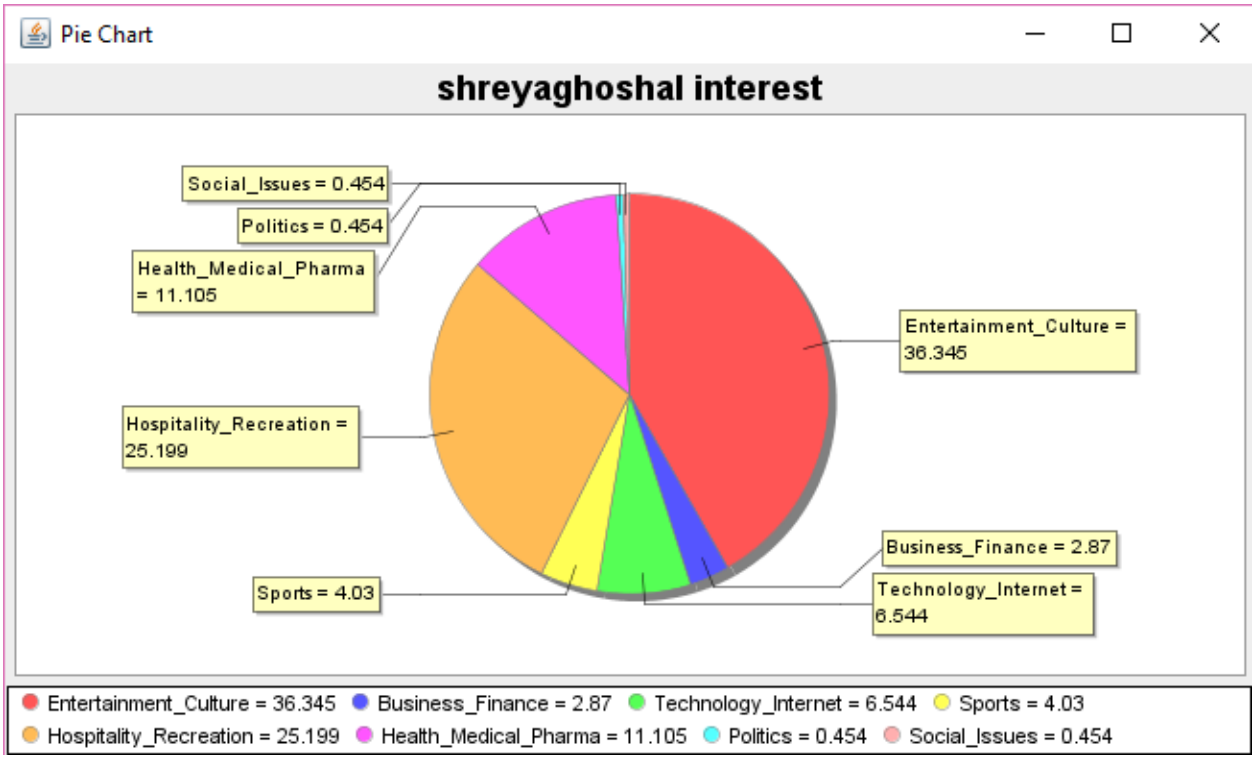


*Figure 4.3*: Show interest pie chart for Shreya Ghoshal

The values shown in the pie chart is in percentage and the entertainment culture topic is having the highest value of the interest that is 36.345% and second interested topic is a hospitality recreation with 25.199 % value of interest.

Below table 4 represents the value in percentage of the interest for certain topics that are listed in table for Narendra Modi, Prime Minister of India, for different period of time.

*Table 4*: Comparison of interested topic for different month for Narendra Modi

| **Topic** / **Month** | **August-15** | **January-16** | **April-16** |
|---|---|---|---|
| Entertainment_Culture | 10.95419584 | 12.0537276 | 8.76078 |
| Business_Finance | 18.22986567 | 14.6631326 | 13.2835 |
| Politics | 24.65910561 | 20.1182752 | 25.4899 |
| Technology_Internet | 2.047844704 | 5.35188855 | 9.44526 |
| Society | 5.413554294 | 4.40708688 | 4.83361 |
| Hospitality_Recreation | 8.248063097 | 11.7906949 | 5.3427 |
| Health_Medical_Pharma | 2.745228883 | 2.29710511 | 1.88765 |
| Education | 3.04303629 | 9.19981396 | 5.4668 |
| Social_Issue | 24.65910561 | 20.1182752 | 25.4899 |

Figure 4.4 shows the histogram representation for the values given in table 4 and in graph it is clearly shown that how the interest of Narendra Modi is changed in different months. We collected tweets for three different months and run algorithm on these different data. For all three months politics and social issue topic have high value of interest. The variation in business finance topic is clearly shown in graph; it is decreased from august-15 to april-16. In august-15 Narendra Modi announced National Handloom Day to mark the 1905 Swadeshi movement and so he did most of the tweet related to Handloom marketing in that time and so in three month business finance topic is peaked in august. With the help of this analysis we can recommend things to user according to his current area of interest as user interest is changed frequently.

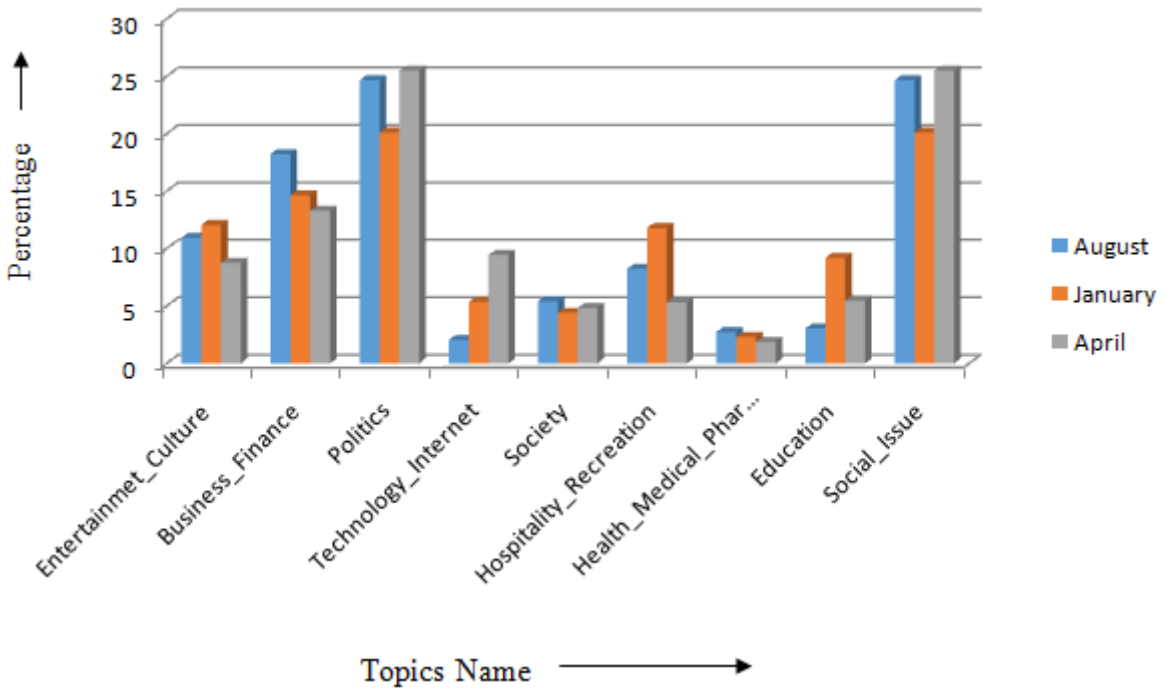


*Figure 4.4*: Shows "interest" histogram for Narendra Modi

### 4.2.2 Location Interest

We have done our experiment over certain cities of India. We collected 2,02,578 tweets of 19 different cities. Figure 4.5 represents the number of tweets done by various cities from 25-02-16 to 18-04-16 time period.

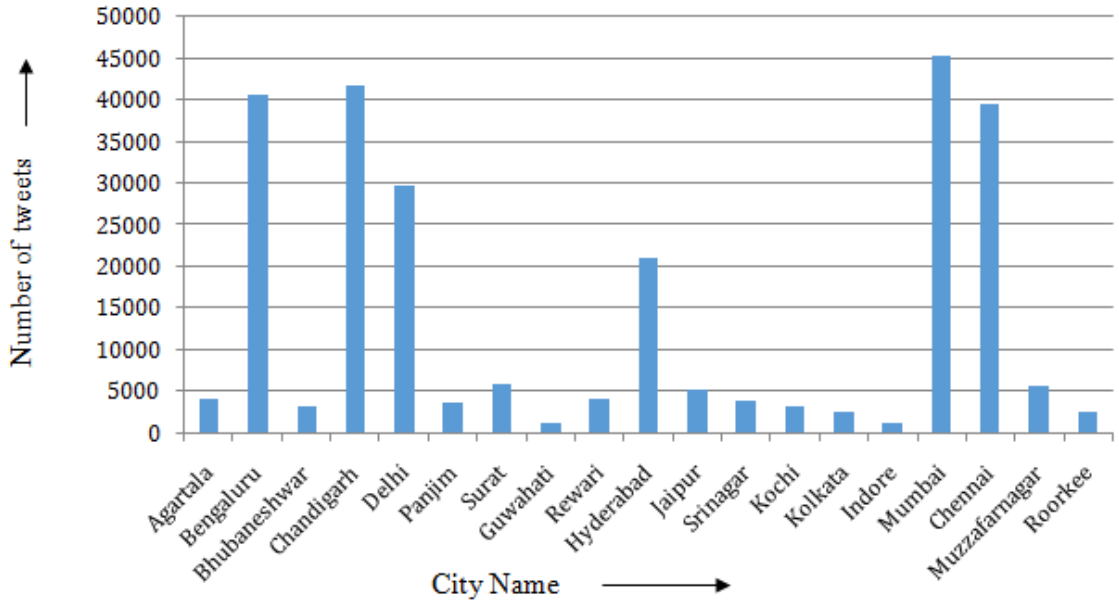


*Figure 4.5*: Number of tweets done by various cities

Below table 5 represents the value in percentage of the interest for certain topics that are listed in table for most of the famous cities of India.

*Table 5*: Comparison data for different cities

| Topic / City | Mumbai | Delhi | Bengaluru | Chandigarh | Hyderabad | Chennai |
|---|---|---|---|---|---|---|
| Entertainment_Culture | 24.2754 | 10.965 | 3.37362 | 9.77781 | 10.5962 | 26.111 |
| Business_Finance | 24.2754 | 12.296 | 50.7963 | 4.65316 | 37.3712 | 16.6319 |
| Politics | 15.5176 | 24.930 | 9.62947 | 10.7892 | 9.57955 | 9.74106 |
| Technology_Internet | 10.2696 | 12.524 | 16.2019 | 3.5699 | 13.7931 | 8.36254 |
| Health_Medical_Pharma | 0.86735 | 4.1926 | 0.54154 | 3.52676 | 1.30405 | 0.85008 |
| Hospitality_Recreation | 5.64081 | 11.307 | 4.07994 | 13.5129 | 6.50522 | 7.855 |
| Social_Issues | 8.00046 | 12.296 | 9.62947 | 26.8084 | 9.57955 | 9.74106 |
| Education | 1.04813 | 2.085 | 1.76921 | 0.5526 | 3.30595 | 1.3003 |
| Sports | 10.0962 | 9.3960 | 3.97375 | 26.8084 | 7.96186 | 19.4044 |
| Other | 0.00893 | 0.0064 | 0.00476 | 0.00009 | 0.0033 | 0.0027 |

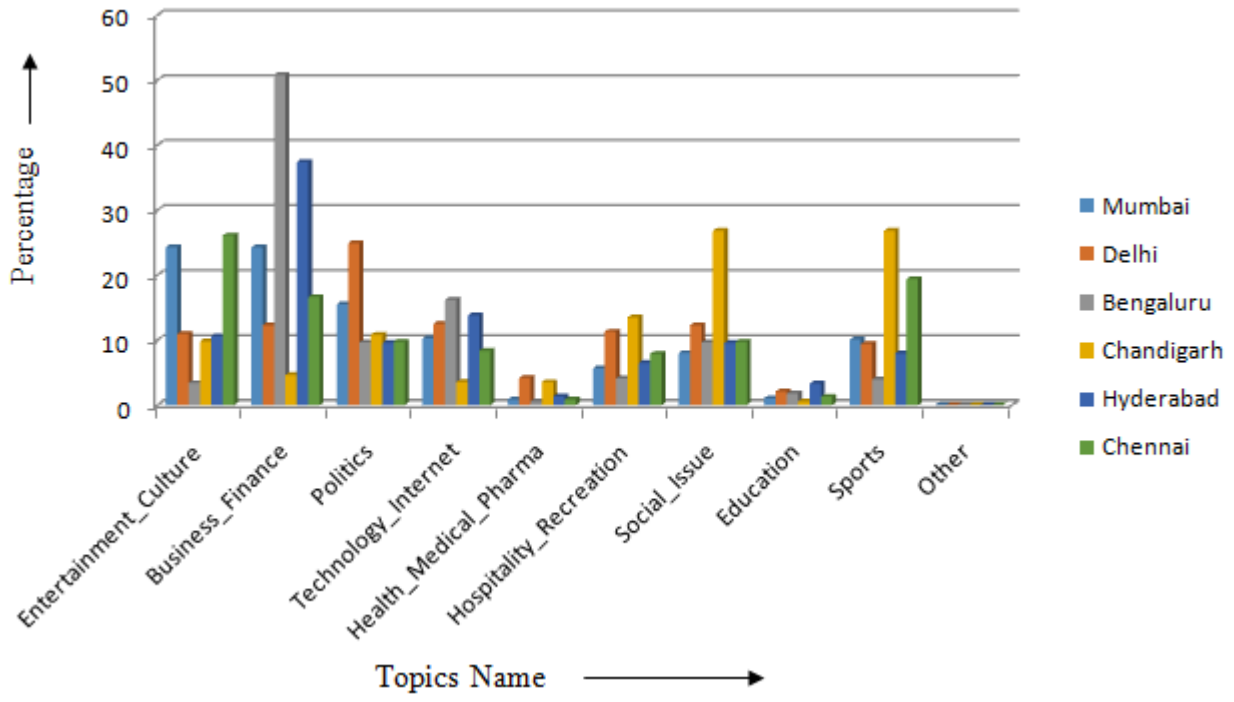


*Figure 4.6*: Shows the comparison of different cities of India

Figure 4.6 represent the graphical representation of the data present in table 5. Bengaluru is one of the cities where most of the multi-national companies are located and most of the business activities take place. The results are showing that among the six cities of India, tweets from Bangalore are highly related to business-finance. Delhi, the capital of India, is a political hub where many politicians and youth that belong to some non- governmental organization (NGO) reside. From the results we infer that among the six cities, tweets from Delhi are mostly related to politics and social issues. Chennai tweets are mostly related to entertainment topic and among six cities Chennai has maximum percentage value ie. 26.11% for entertainment topic.

It is useful to have such data, information because it provides us the trend of users at certain locations. Based on the trends of the twitter data, new products may be launched in certain locations.


*Figure 4.7*: Represent the major interested topic on marked location

Figure 4.7 represents the major interested topic for the cities which we have shown in comparison graph above in figure 4.6 as well as the cities which we do not show in above comparison graph. The value at markers represent topic name and description of topic is given below in table 6.

*Table 6*: Represent topic name marked at particular city

| **City Name** | **Topic Marked** |
|---|---|
| Mumbai | E- Entertainment_Culture |
| Delhi | P- Politics |
| Bengaluru | B- Business_Finance |
| Chandigarh | S- Sports |
| Hyderabad | B- Business_Finance |
| Chennai | S- Sports |
| Kolkata | B- Business_Finance |
| Kochi | T- Technology_Internet |
| Jaipur | H- Hospitality_Recreation |
| Panjim | T- Technology_Internet |
| Indore | B- Business_Finance |
| Surat | T- Technology_Internet |
| Guwahati | E- Entertainment_Culture |
| Roorkee | A- Social_Issue |
| Bhubaneshwar | H- Hospitality_Recreation |
| Muzzafarnagar | B- Business_Finance |
| Rewari | P- Politics |
| Agartala | S- Sports |
| Srinagar | A-Social_Issue |

### 4.2.3 Comparison of countries data

Below table 6 represents the value in percentage of the interest for certain topics that are listed in a table for two countries that are India and America. We have obtained these values by collecting the tweets of different twitter users of India and America and here we try to get the overall interest of Indian and American users and our analysis shows that most of the tweet done by Indian users are related to politics and social issues with 18.291% value of interest and then other interested topic is hospitality recreation with 14.393% value of interest.

The tweets done by American users are mostly related to hospitality recreation with 23.986% value of interest and then second major interested topic is entertainment with 15.9% value of interest. All the tweets of American users are belonged to some topic so percentage for the category other is blank in this case. Figure 4.7 shows the comparison of Indian and American users graphically based on the tweets collected for both the countries.

*Table 7*: Comparison data for countries, India and America

| Topic \ Country | India | America |
|---|---|---|
| Entertainment_Culture | 10.625 | 15.9 |
| Business_Finance | 9.5233 | 8.3479 |
| Politics | 18.291 | 5.8073 |
| Technology_Internet | 5.5157 | 14.754 |
| Sports | 12.85 | 14.464 |
| Hospitality_Recreation | 14.393 | 23.986 |
| Health_Medical_Pharma | 5.6224 | 7.3929 |
| Education | 4.8878 | 3.5413 |
| Social_Issues | 18.291 | 5.8073 |
| Other | 6.90E-05 | |

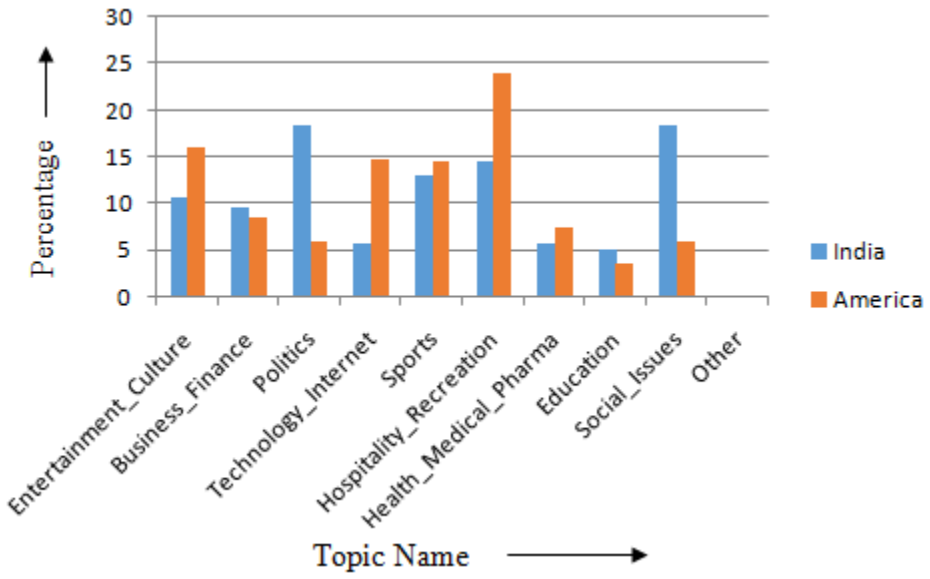


*Figure 4.8*: Shows the comparison between Indian and American users

### 4.2.4 Comparison of sentiment analysis algorithm

Initially, we are using bag of word approach which contains some text file having human annoted codes (sentiment score) for several words and final sentiment score is computed on the

basis of this score [3]. This bag of word approach is done using senti strength jar file which has some predefined function for computing sentiments. This approach is appropriate for the large text file but tweets are very short informal text and many user uses acronym and emoticon in tweet to show their emotions because of restriction in tweet length. For this short text analysis we are using Stanford CoreNLP sentiment method which forms a labeled tree structure. Detailed working of this method is explained in chapter 3. The drawback of this method, it is not considering acronym and emoticon value properly. If we calculate sentiment score for happy face and sad face using this approach than this method provides neutral result for both cases. To overcome this problem we generated emoticon and acronym text files and before calling sentiment procedure we first check the sentiment of text in these text file. If found then consider the sentiment score for this which is not considered in previous approach. We have tested this approach on some data and result is shown below in figure 4.9.

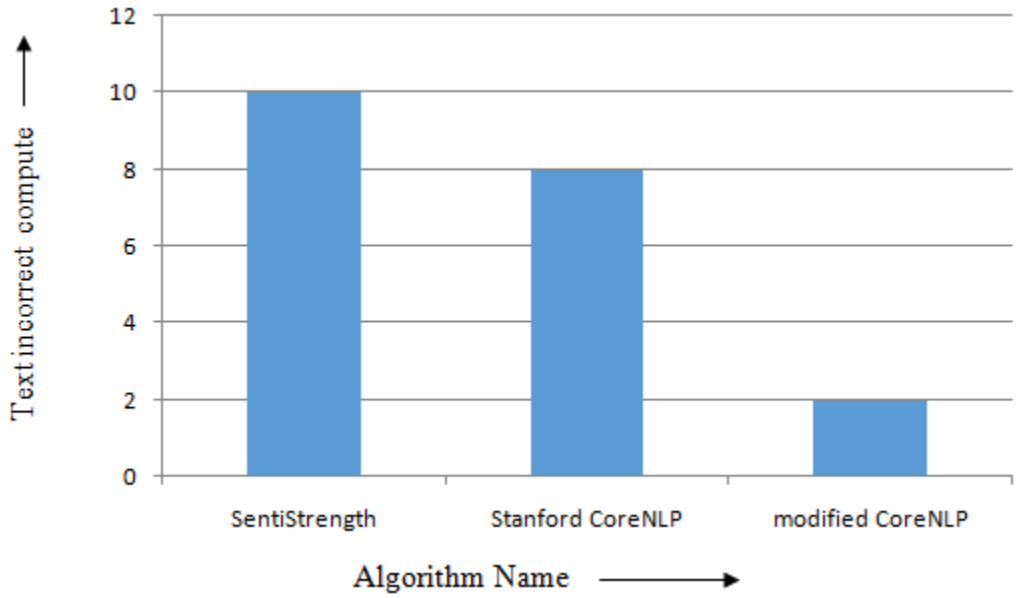


*Figure 4.9*: Comparison of sentiment algorithm used

Figure 4.9 represents the comparison of sentiment algorithm we used. Here, y-axis represents the number for which these algorithm work incorrectly. We have tested these algorithm on 50 different short sentences. We have found that sentiStrength algorithm incorrectly work for 10 sentences, Stanford CoreNLP algorithm incorrectly work for 8 sentences and modified CoreNLP algorithm incorrectly work for 2 sentences.

CHAPTER 5

# Conclusion And Future Work

Since now, social media become a part of human life. So analysis of data produced by such sites and get something knowledgeable for humanity is also useful. This data help us to solve several problem. In this thesis, we deal with a problem which analyse the user interest. For analysis, we have done sentiment analysis and categorization of tweets. Based on these results we have shown the user's interest and the current trending topic going on certain locations. We have taken million of tweets and using this we infer that this research helps in product launching. This work helps in betterment of any recommendation system. We also compare sentiment analysis algorithm that we are using and try to use accurate sentiment algorithm in our thesis. In future we extend our work by improving the algorithm used for classification of tweets.

# BIBLIOGRAPHY


[1] R. Prabowo1, and, M.Thelwall, "Sentiment Analysis: A Combined Approach," Published in Journal of Informetrics , Vol. 3(2), pp 143-157, 2009.

[2] A. Agarwal, B.Xie, I.Vovsha, O.Rambow, and, R.Passonneau, "Sentiment Analysis of Twitter Data," Proceeding LSM'11 Workshop on Languages in Social Media. Association for Computational Linguistics, pp 30-38, 2011.

[3] M.Thelwall, "Heart and Soul: Sentiment Strength detection in the Social Web with SentiStrength," Proceedings of the CyberEmotions, pp 1-14, 2013.

[4] D. F. Gurini, F. Gasparetti, A. Micarelli, and, G. Sansonetti, "A Sentiment-Based Approach to Twitter User Recommendation," Published in 5th ACM RecSys workshop on Recommender Systems and the social web, June 2013.

[5] V. Agarwal, and, K. K. Bharadwaj, "A collaborative filtering framework for friends recommendation in social networks based on interaction intensity and adaptive user similarity," Published in Journal Social Network Analysis and Mining, Vol. 3, pp 359-379, 2013.

[6] F. Abel, Q. Gao, G.Houben, and, K. Tao, "Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web," Published in The Semantic Web: Research and Applications. Springer Berlin Heidelberg, pp 375-389, 2011.

[7] A.Boutet,H. Kim, and, E. Yoneki, "What's in Twitter, I know what parties are popular and who you are supporting now!," Published in Journal Social Network Analysis and Mining, Vol. 3(4), pp 1379-1391, 2013.

[8] A. Java, X. Song, T. Finin, and, B. Tseng, "Why we Twitter: Understanding Microblogging Usage and Communities," Proceedings of 9th WebKDD and 1st SNA-KDD Workshop, SanJose, California, USA, pp 56-65, August 2007.

[9] M.Pennacchiotti, and, A.Popescu, "A Machine Learning Approach to Twitter User Classification," Proceedings of the Fifth ICWSM, pp 281-288, 2011.

[10] K. Lee, D. Palsetia, R. Narayanan, Md. Mostofa Ali Patwary, A. Agrawal, and, A. Choudhary, "Twitter Trending Topic Classification," Published in 11th IEEE International Conference on Data Mining Workshops, pp 251-258, December2011.

[11] R. Socher , A. Perelygin , J. Y. Wu , J. Chuang , C. D. Manning , A. Y. Ng , and, C. Potts," Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," Proceedings of the conference on empirical methods in natural language processing, Citeseer, pp 1631-1642, 2013.

[12] S. Kiritchenko, X. Zhu, and, S. M. Mohammad, “Sentiment Analysis of Short Informal Texts,” Published in Journal of Artificial Intelligence Research, Vol 50(1), pp 723-762, USA, May 2014.

[13] Manning, D. Christopher, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, ”The Stanford CoreNLP Natural Language Processing Toolkit,” Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60, 2014.

[14] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features,” Proceedings of the 10th European Conference on Machine Learning, London, pp 137-142, 1998.

[15] L. Huges and, L. Palen, "Twitter Adoption and Use in Mass Convergence and Emergency Events," Proceedings of International Journal of Emergency Management, Vol 6(3-4), pp 248-260, May 2009.

[16] Y. Sharma, D. Bhatia and, V.K. Choudhary,"TwiBiNG: A Bipartite News Generator Using Twitter," Proceedings of the SNOW 2014 Data Challenge co-located with 23rd International World Wide Web Conference (WWW 2014), pp 70-76, Seoul, Korea, April 8, 2014

[17] C. C. Aggarwal and, K. Subbian, "Event Detection in Social Streams," Proceedings of *SDM*, Vol 12, pp 624-635, 2012.

[18] H. Abdelhaq, C. Sengstock and, M.Gertxz, "EvenTweet:Online Localized Event Detection from Twitter," Proceedings of the VLDB Endowment,Vol. 6(12), 2014.

[19] G.Ifrim, B. Shi and, I. Brigadir. "Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering," Proceedings of SNOW WWW Workshop, 2014.

[20] X. Wang, F. Zhu, J. Jiang and, S. Li, "Real Time Event Detection in Twitter," Proceedings of Web-Age Information Management, pp 502-513, Springer Berlin Heidelberg, 2013.

[21] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and, I. M. Welpe, “Predicting Elections with Twitter: What 140 characters reveal about political sentiment,” Proceeding of the Fourth International AAAI Conference on Weblogs and Social Media, 2010.