A DEMOGRAPHY BASED ANALYSIS OF USERS' SENTIMENTS ON TWITTER DATA

A DISSERTATION

Submitted in partial fulfillment of the requirements for the award of the degree

of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

GEETA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE -247 667 (INDIA)
MAY, 2016

A DEMOGRAPHY BASED ANALYSIS OF USERS' SENTIMENTS ON TWITTER DATA

A DISSERTATION

Submitted in partial fulfillment of the
requirements for the award of the degree

of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

GEETA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE -247 667 (INDIA)
MAY, 2016

CANDIDATE'S DECLARATION

I hereby declare that the work, which is being presented in the dissertation entitled **''A demography based analysis of users' sentiments on twitter data''** towards the partial fulfillment of the requirement for the award of the degree of **Master of Technology** in **Computer Science and Engineering** submitted in the Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand (India) is an authentic record of my own work carried out during the period from July 2015 to April 2016, under the guidance of **Dr. Rajdeep Niyogi, Associate Professor,** Department of Computer Science and Engineering, IIT 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:

Place: Roorkee **(Geeta)**

CERTIFICATE

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

Date:

Place: Roorkee **(Dr. Rajdeep Niyogi)**

Associate Professor

Department of Computer Science and Engineering

IIT Roorkee

ACKNOWLEDGEMENTS

I feel much honoured in presenting this seminar report in such an authenticable form of sheer endurance and continual efforts of inspiring excellence from various coordinating factor of cooperation and sincere efforts drawn from all sources of knowledge.

I express deep gratitude for en enthusiasm and valuable suggestions that I got from my guide Dr. Rajdeep Niyogi, Department of Computer Science & Engineering, Indian Institute of Technology Roorkee, for successful completion of this project. His integrity and commitment has inspire me.

I heartily thank professors and technical staff of Computer Science & Engineering, IIT Roorkee, for helping me in exploring the possibility of carrying out the experiments in their laboratory.

Finally, I would like to acknowledge with gratitude, the support and love of my family and friends. They all kept me going.

Geeta

ABSTRACT

In recent times the popularity of social media has gone up greatly. Due to the great number of people using these platforms and going vocal with their thoughts these can be used to determine public opinion. In this report, we are extracting this opinion of people by analyzing the tweets collected from twitter on major events like T20 World Cup, Paris Attack, Oscar, Olympics, Formula 1 championship etc. Here we have used demographic analysis. We first analyze the opinion of users and then calculate the sentiments of users on different events. In this way, we determine how users' opinion and their positive and negative sentiments differ demographically. With the help of Sentiment Analysis techniques we analyze the demographic behavior of users. It consists of three major modules: a data collection and preprocessing module, apply Opinion Mining, apply Lexicon based Approach for Sentiment analysis. We have performed this analysis on millions of twitter users residing in different locations and have demonstrated the findings using bar charts and pie charts.

Keywords— sentiment analysis; social media; demographics; opinion mining.

.

Table of Contents

1. INTRODUCTION....................1

1.1 Overview....................1

1.2 Motivation....................2

1.3 Problem Statement....................3

1.4 Organization of Thesis....................3

2. LITERATURE REVIEW....................4

2.1 Gaps Identified....................7

2.2 Objectives....................7

3. PROPOSED METHODOLOGY....................8

3.1 Sentiment Analysis....................8

3.2 Demographic analysis....................8

3.3 Proposed Approach....................9

3.4 Twitter Demography....................10

3.5 Work Flow....................11

3.6 Data Collection....................12

3.7 Geolocation....................12

3.8 Opinion Mining and Sentiment Analysis....................13

3.9 Proposed Method....................14

4. IMPLEMENTATION DETAILS....................15

4.1 Latitude and Longitude....................16

4.2 Download Data....................16

4.2.1 Web Scraping....................16

4.2.2 Setup Twitter API Credentials....................16

4.3 Extract Tweets....................18

4.4 Preprocessing....................18

4.5 Opinion Mining....................20

4.6 Determine the Sentiments through AFFIN file..20

5. RESULTS AND DISCUSSION...22

5.1 Results..22

5.2 Discussion...28

5.3 Limitations..29

6. CONCLUSION AND FUTURE WORK..30

REFERENCES **31**

List of Tables

Table 3.1 Social Media Data for Different Countries..10

Table 5.1 Opinions of users in Percentage..22

Table 5.2 Positive Sentiments in Percentage..25

Table 5.3 Negative Sentiment in Percentage...25

List of Figures

Figure 3.1 Workflow of Proposed Solution..11

Figure 4.1 Flow Chart of Complete Implementation Details..15

Figure 4.2 Latitude and longitude of India...16

Figure 4.3 Twitter API Credentials...17

Figure 4.4 Unprocessed Twitter Feeds...17

Figure 4.5 Refinements of Tweets...18

Figure 4.6 Pseudo Code for processing Steps..19

Figure 4.7 Calculate Positive and Negative Polarity..21

Figure 5.1 Opinions of Five Countries about Five Events..23

Figure 5.2 Percentage of each Event given by each Country..25

Figure 5.3 Positive and Negative Sentiment Scores of Users...28

CHAPTER 1

INTRODUCTION

1.1 Overview

In today’s fast moving world where people have lost personal touch online social networks serve as a platform which people are actively using in order to connect, communicate and express themselves to others. Active users regularly post content and express their thoughts on a plethora of issues ranging from politics to sports. Any application where data is involved can serve as a great source of knowledge. In case of online social networks the content posted by the users serves as the data that can be analyzed to mine knowledge. Sentiment analysis and opinion mining are two data mining tasks that are usually performed on social network data to extract knowledge [8].

Twitter is an on-line social networking service and a micro blogging service. Twitter allows its users to post and read short messages of not more than 140 characters called as "tweets". There are about 320 million monthly and 100 million daily active users of twitter in the world[23].

Demographic analysis of the content shared by twitter users can be useful for social scientists, marketers and policy makers. In this report, we first look at the demographics of the twitter users and explore if and how users from different countries vary. We have picked some important events and then we analyze the content pertaining to users of several countries related to these events and try to find out whether their opinions are demographically based or not. After that we do sentiment analysis to know their positive or negative views on those events.

Understanding users' sentiments has its applications in many domains. For instance, marketing department can benefit if data extracted from social media is mined to determine the reaction of people to any product or service. Similarly social scientists can use social network data to study human behavior and reaction to various different events. An analyst must be aware of existing sentiment differences.

In this report, we try to figure out the opinion of users of different countries in regard to some key events. First, we take up users for analysis such that they are geographically distributed. Post that, we collect tweets using twitter public API based on longitude and latitude of countries in python, i.e. available for general public for free. We have collected tweets of users from five different countries- United States, India, Brazil, Australia and France pertaining to five key events- T20 World Cup, Paris Attack, Oscar, Formula 1 championship and Olympics. After that, we perform a thorough analysis of the data collected using mining techniques and examine whether sentiments and opinions are demographically based or not.

The basic aim of sentiment analysis is to determine whether a user has positive or negative sentiments for the particular brand, event or product. Sentiment analysis uses natural language processing with artificial intelligence capability and text analytics to analyze the collected data to determine the same. Sentiment Analysis is performed on three basic nodes: in which document is considered, where sentence is considered, and here aspect is considered [1]. The basic approches used for determining sentiments can be mainly divided into three categories: machine learning techniques and lexicon based techniques and hybrid techniques[1]. The lexicon based approach relies on a collection of known and preprocessed sentiment terms, machine learning approaches rely on various algorithms for analysis and the hybrid approaches use a combination of these two approaches [1]. In this report, we are following a lexicon based approach to demonstrate how users' sentiments expressed on twitter regarding five different events vary demographically in terms of overall sentiment score.

1.2 Motivation:

The popularity as well as the widespread use of OSNs by people all around the world has influenced researchers all around the world and researchers are increasingly using this data in order to study social behavior and relationships. Also the content shared as well as the views expressed by people on OSNs influence the opinions of other people using them. How this influence changes general opinion as well as behavior is also a popular research area. In the past not much heed has been paid to the use of OSN data in order to determine demographic information related to users so as to understand behaviors and attitude. Such type of research is pivotal for social and behavior scientists. We have combined these two approaches to understand whether users’ opinion or sentiments vary demographically or not by analyzing the twitter data of users belonging to different countries on different subjects.

1.3 Problem Statement:

The demographic information of a population can be useful for social scientists, marketers, and policy makers. The problem i.e. stated in this report is: *Are there any demographic difference about opinions among different countries' users for a certain event that happens on a certain location ?*

1.4 Organization of the Report:

First chapter provides overview of the online social networking sites. A basic information about Twitter and process of our approach is discussed. Later part of the section gives the motivation and need behind taking up this area of research. Finally it establishes the problem statement.

Second chapter is literature review section which talks about the advancement in the research area chosen and the research gaps.

Third chapter gives us the framework of the project and work done. It explain combined approach that is chosen by us and work flow of our process is described.

Fourth chapter provides implementation details of our approach. It also shows how we collect the data and how we find out the opinions and sentiments of users.

Fifth chapter display the result. Results for each country and each event will be shown and comparison of result variation will also be discussed. After that some limitations of this work will be discussed

Sixth chapter gives conclusion of the research and we will discuss about the future work. In what way research can be extended so that obtained results can be improved further.

CHAPTER 2

LITERATURE REVIEW

Lots of research has been done on Sentiment Analysis and Opinion Mining and several different methods are proposed for this purpose. In many research papers, Machine learning techniques, Lexicon based approach etc. for sentiment analysis has been used for several application. A few other studies have examined the demographics of social network users. Here are some literature work that is related to my work :

Agarwal and Xie [3] introduces POS-specific prior polarity feature for sentiment analysis, it uses tree kernel approach. Three experimental sets are framed: feature based model uses hundred features. Accuracy remains same to that of thousand features. First phase in Kernel tree based model is to tokenization of tweets into a tree. It is done by differentiating punctuations mark, emotion and other features as per suitability. It also determines word polarity by looking into word net dictionary.

Arora, Li and Neville [1] presents in their published research work about sentiment analysis on twitter data to look into the received response on smart phone brands and functionalities of operating system used in them. This is done through Lexicon based sentiment analysis. This technique incorporates three steps. First one is data collection and cleansing, second one is sentiment classification and last one is determination of overall sentiment score.

Sarlan, Nadam and Basri [7] present paper on sentiment analysis .They have used machine learning based approach which worked in conjunction with natural language processing techniques. Natural language processing technique (NLP), support vector machine (SVM), artificial neural network technology(ANN) , and case based reasoning (CBR) , all of them uses polarity assignment methodologies of sentiment analysis. A process is created that distributes the sentiment into positive sentiments and other is negative sentiments. Limitation of Django and LAMP creates hindrance in making of program as web application.

Neethu and Rajasree [6] paper analyses the post in twitter regarding electronic products or items eg palmtops , ipads etc using machine learning. Classifying tweets is important and hence a new feature vector is created which classifies as positive , neagative.It also extract users view about items.two basic techniques exist for deterring sentiment from texts. Symbolic and machine learning constitutes them.

Naive Bayes, Maximum Entropy and Ensemble classifier and Support Vector Machine ahev ebeen used for classification and their respective performance is compared. MATLAB simulator has been used for above mentioned classifier. Proper Testing has been performed. IT has came out that Naïve Bayes has better precision when, is compared to all of them mentioned above.

Shulong Tan et al.[18] has dealt with the problem which referred to public sentiment variation and finding its those possibilities which caused these variations. two Latent Dirichlet Allocation (LDA) based models, Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA have been proposed for solving the problem.

Hassan, Abbasi and Zeng [19] proposed sentiment analysis on a text analytics framework for twitter. An elaborate bootstrapping ensemble has been used to deal with imbalance ,sparsity and representational richness of class. Experiments shows that results obtained from this approach are more reliable and balanced as compare to other methods and process. Bootstrapping Ensemble Framework (BPEF) is a twin stride process: expansion stage and contraction stage.

Gao, Berendt and Vanschoren [20] have shown the difference in sentiment through privacy level and demographic factors. Chats of facebook users and posts in various languages are used as dataset with their parameters and features. An two algorithms approach has been used, that generalizes single attribute testing by referring to subgroup-discovery paradigm.

Sloan and Morgan [4] has shown the demographic characteristics of twitter users in order to show the demographic differences between those user who uses location services provided to them and those who do not use them. And those who geotag their event and those who do not geotag their tweets. Two types of dataset is used for this purpose, one which dwell on enabling geoservices and the otherwhihc greatly focuses on tweet geotagging. Twitter user's behavious has been investigated considering their age, class (economic and social) , gender and language.

Oktay et. al. [12] propose an algorithm for age estimation which uses people's first names. A Bayesian generative model has been used to estimate age using their respective first name and their particular ethnicity using their respective last name. They find that different demographic groups both in term of age and ethnicity have different usage patterns on the platform in terms of topical conversation and the time in the day to use the platform.

Murthy, Gross and Pensavalle [5] shows us the investigation on intersection between gender ,place, race and ethnicity amongst Twitter users particularly from America. The intensity of activity done by users on Twitter is measured by two approaches: Power Law Behavior Method and Inter Tweet Interval Method.

Misolve et. al. [2] investigated demographics considering countries and used first their name as a proxy to know their gender and last name to know their race or ethnicity. These trends explore few of the basic level demographic changes occurred in particularly at American usage of Twitter. Self reported location has been used by them and information gained from their names in Twitter profile to explore demographics of users along their geographic region, ethnicity and gender.

Sloan et. al. [9] specifies, designs and critically evaluates two tools for the automated identification of demographic data (age, occupation and social class) from the profile description of twitter users in the United Kindom (UK). Meta-data data routinely collected through the collaborative social media observatory. Occupation detection algorithm is used for this purpose.

McCormick et. al. [11] creates a toolkit for researches produced in social science and interested in using twitter data for behavioral examination and attributes. Some of the new approaches for analysis of data from social media, their extraction and their processing offered. Case control sampling framework is used for data extraction.

Mitchell et.al. [13] shows a elucidated investigation of mapping between real-time expressions of persons made particularly in United States and a wide range of geographic, demographic, emotional, and health characteristics. Information about how geographic place correlates with and potentially influences societal level of happiness. Urban areas are defined by the 2010 US census bureau's MAF/TIGER database. To measure sentiment from the words collected, we use the language assessment by Mechanical Turk (LabMT) wordlist.

2.1 Gaps Identified

Though a lot of work has been done but based on the literature survey carried out and presented above , following gaps are identified:

In [5], the investigation on intersection between gender ,place, race and ethnicity amongst Twitter users particularly from America. But they only consider who is tweeting. There is a need in the social sciences and beyond to understand not only who is tweeting but how intensely different groups tweet.

In [12], an algorithm is proposed for age estimation which uses people's first names but they were not consider the thing that if different age groups post about different topics, than in what frequency different age groups talks about different topics.

2.2 Objectives of the present study

The proposed work has the following objectives and contributions in the field of Twitter data analysis :

We investigate whether the comparison is proper for the event. Data for those event has been taken from twitter. The five events considered in this research are: T20 World Cup, Paris Attack, Oscar, Formula 1 championship and Olympics which are held on India, France, US, Australia and Brazil respectively. Our objective of this study is to analyze the demographic behavior of users with the help of sentiment analysis techniques.

It consists of three major modules:

- A Data Collection and Preprocessing module
- Apply Opinion Mining
- Apply Lexicon Based Approach for Sentiment Analysis

CHAPTER 3

PROPOSED METHODOLOGY

Online Social Networks (OSNs) are today very prevalent in our lives and are used my many users to connect, communicate, and share content. Due to the wide spread use of OSNs, huge amount of data can be collected from them about the users as well as the way they communicate and this data can be analyzed to determine a lot about the human society. However, the sensitivity of the data as personal information is involved leads to privacy constraints. And thus such data cannot be released for analysis. Hence, data from most OSNs cannot be used for analysis as it is. In this case twitter is an exception as more than 90% of twitter users have set their profile visibility as public and also their communication history is public and can be viewed. This allows researchers to gather a lot of information from the OSN. And therefore in this work we have harnessed the power of data offered by twitter to study the communications of a large fraction of the population.

As described in the literature review a lot of research has been done on twitter data. This includes analysis like sentiment analysis and demographic analysis.

3.1 Sentiment Analysis

A lot of work has been done under this area as discussed in chapter 2. Sentiment analysis is the analysis of data in order to establish the sentiments of a particular set of population in regard to some particular product, service or event. It can be used to know the general views in regard to a event, product or service. Here in this work we will be using sentiment analysis in order to analyze the sentiments of the users in relation to a set of events. This is one application, sentiment analysis can also be used to establish the views of public on a new phone or a particular brand. Several techniques has been used for this purpose for example Lexicon Based Approach, Machine Learning Techniques, Hybrid Approach, Tree-Kernel Based Approach etc.

3.2 Demographic Analysis

The purpose of demographic analysis is to analyze the way the population of a particular area behaves. It is the study of the behavior of a particular area’s population by the analysis of data collected regarding their communicating patterns and their content. Here we have used twitter data i.e. tweets in order to study the population of a particular area. In previous researches, some work is done to know what if users show there geographic information or not. In some other cases demographic analysis is used to know the gender of the users, the common characteristics of group of users of twitter.

3.3 Proposed Approach

In the past a lot of work has been done to perform sentiment analysis as well as demographic analysis. Here in this work we have combined both of these analyses. The popularity as well as the widespread use of OSNs by people has influenced researchers all around the world and researchers are increasingly using this data in order to study social behavior and relationships. Also the content shared as well as the views expressed by people on OSNs influence the opinions of other people using them. How this influence changes general opinion as well as behavior is also a popular research area. In the past not much heed has been paid to the use of OSN data in order to determine demographic information related to users so as to understand behaviors and attitude. Such type of research is pivotal for social and behavior scientists. The demographic information of a population can be useful for social scientists, marketers, and policy makers.

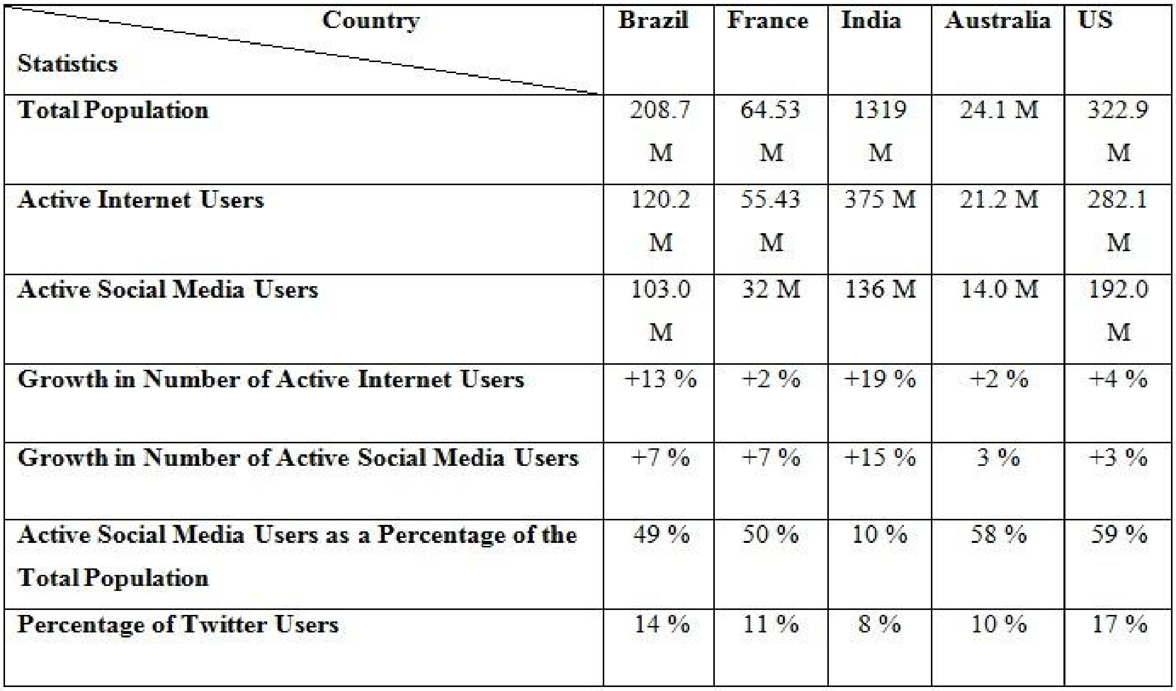
Most of the work done in the past is related to performing sentiment analysis in order to determine the sentiments of users about a product or brand or on finding demographic information of the users. However, these works do not study how users' opinion is affected by demographics. Or in other words we can say that these works do not show whether user opinions are demographic based or not. Our approach is different from others as we use social network data to determine the opinion of users of different countries and further analyze the positive or negative sentiments of these users belonging to different countries.

If we only perform sentiment analysis then we will only know about sentiments of users which may be positive or negative in relation to any event, product or service. We will not be getting the information whether those positive or negative views based on the factor where that event was held. If we only find the demographic characteristics of a country, than we don't get the knowledge about how different the characteristics of different countries are. So here we have combined these two approaches to understand whether users' opinion or sentiments vary demographically or not by analyzing the twitter data of users belonging to different countries on different subjects.

3.4 Twitter Demography

Out of a total population of 2.307 billion active OSN users around 320 million users are active on twitter. The daily count of users seen active by twitter is around 100 million and these users send around 5 million tweets in a single day [23]. As evident from these numbers, huge amount of data can be collected from twitter. Out of the total 320 million twitter users 65 million belong to the US. In Table 1, we have described the comparison of demographic data of Brazil, India, France, US and Australia. In which we shows the statistics like total population of a country, how many users that are active on social media, total number of twitter users in that country etc. We can see that the US has the highest percentage of twitter users and it is the number 1 country in the world with most active twitter users with 120 million users. Brazil has approximately 40 million active twitter users. With 17.5 million active users, India is also in the top ten most active countries on twitter.

Table 3.1 Social Media Data for Different Countries[24]



Country Statistics	Brazil	France	India	Australia	US
Total Population	208.7 M	64.53 M	1319 M	24.1 M	322.9 M
Active Internet Users	120.2 M	55.43 M	375 M	21.2 M	282.1 M
Active Social Media Users	103.0 M	32 M	136 M	14.0 M	192.0 M
Growth in Number of Active Internet Users	+13 %	+2 %	+19 %	+2 %	+4 %
Growth in Number of Active Social Media Users	+7 %	+7 %	+15 %	3 %	+3 %
Active Social Media Users as a Percentage of the Total Population	49 %	50 %	10 %	58 %	59 %
Percentage of Twitter Users	14 %	11 %	8 %	10 %	17 %

3.5 Work Flow

In this section we describe the complete workflow of our framework. We will discuss all the steps one by one that will be used in our research. The below figure 1 shows the basic flow diagram of our method. Here in this work we have chosen 5 key events about which sentiment analysis will be performed for the users of 5 different countries viz. India, US, Brazil, France and Australia. The first step involved in any data analysis is collecting the data. So first, we have collected the tweets of the user from twitter API based on their location. Here, the locations are the countries that we chosen and we collect the data based on the latitude and longitude of those countries. Second, we have performed some preprocessing on the tweets. Opinion Mining has been done on the tweets collected in order to know how different the views of users on an event are. What percentage of users share the same opinion. Sentiment analysis approach is used to know that what are users sentiments, what they think about a particular event, their views either positive or negative. Post performing sentiment analysis we use the determined opinions of the users in order to perform comparison among users of different countries i.e. demographic analysis is performed. Finally we have summarized the observations made during demographic analysis using bar charts and pie charts.

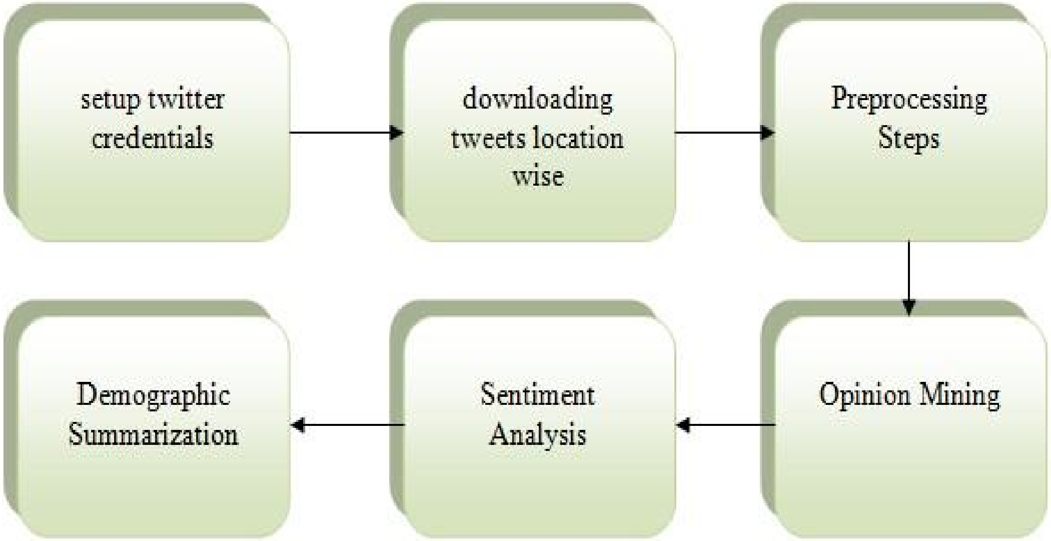

setup twitter credentials
downloading tweets location wise
Preprocessing Steps
Opinion Mining
Sentiment Analysis
Demographic Summarization

Fig. 3.1. Workflow of proposed solution

3.6 Data Collection

In order to have data pertaining to different countries we have used the geocode locations as in latitude and longitude, in Twitter's Stream API, that requests tweets posted by a particular country and by making use of latitude and longitude points. Mentioned method has been actively used in the past as well for collecting and then studying twitter data geographically. Twitter data is collected for five events for users of five different countries. For this purpose, first of all we find out the latitude and longitude coordinates of a country using latlong.net [21]. In addition to the data content, downloaded tweets also provide us other information such as the number of followers, location, number of re-tweets, count and hash-tags. All these tweets are collected using Twitter API [22].

3.7 Geolocation

Twitter users have the option to enable or disable location services on their account. By default location services are off and a user can enable them if he or she wishes to geotag his/her tweets. Once the users enable location services then there exact longitude and latitude points can be determined from their tweets. From the perspective of social analysts and scientists this location data is immensely valuable as it allows them to understand the geographic context at the time of data creation itself. Although the proportion of geotagged tweets is small but it still helps us to know where a person is when they publish the tweet.

The geo-tagging feature in the Twitter API allows attaching location to a tweet. This feature enables a more meaningful experience for the users by making tweets more contextual. To determine tweets pertaining to the same geographical space the API does not provide a “near” search operator, but instead provides a more exact way to restrict your query to a geographical area by making use of the geo-code parameter. The geo-code parameter represents the geographical information in the form of a triplet defining the location in terms of “latitude, longitude, radius”, for instance, “37.781157,-122.398720, 1mi” defines a particular geographical location. When conducting geo searches, the search API will first attempt to find tweets which have lat/long within the queried geo-code, and in case this attempt is not successful, then it will try to find tweets posted by users whose profile location can be reverse geo-coded into a latitude and longitude that lies within the queried geo-code, meaning that the API can also return tweets as response to the query which do not include lat/long information.

3.8 Opinion Mining and Sentiment Analysis

With the explosive growth of *social media*, individuals and organizations are increasingly using public opinions in these media for their decision making. What public think about particular brand, event, political issue etc. is very important. How much positive response they give can be determine by analyze their sentiments. For this purpose we have used Lexicon Based Sentiment Analysis Approach to understand the personal opinion of a user about an event. In lexicon based analysis we look for opinion words in the data viz. tweets in order to determine the whether a user's opinion is positive or negative. Opinions words symbolize sentiments of the author. It may be positive or negative sentiment. The collection of opinion words that is referred to find sentiment positivity (positive/negative) is called the opinion lexicon. This methodologies of using such words to find opinion characteristics and pattern is called the lexicon based approach to sentiment analysis.

Much has been described about sentiment analysis. Majorly it is based on following steps.

- Data collection and its refinement
- Sentiment characterization and classification
- The determination of sentiment score finally

Analysis described in this research Twitter data has been reffered to determine users' opinions about various events i.e. T20 World Cup, Paris Attack, Oscar, Formula 1 championship and Olympics. These users belong to five different countries viz. India, France, US, Australia and Brazil. The data has been determine when Twitter API is connected by using Python libraries.

Sentiment analysis has always been heavily dependent on different phrases and various words. Which can be categorized into positive or negative sentiment i.e. the opinion words are used to perform lexicon based sentiment analysis.

3.10 Proposed Method

Following is the step by step process of our proposed solution:

1) Find out the latitude and longitude of a Country.
2) Collect the data that depends on the location parameter in Twitter's Streams API.
3) The data we collect contains information like users' screen name, full name, tweet text, tweet id, followers, re-tweets, location etc. Out of which we extract only tweets.
4) Apply preprocessing steps by removing slang words, stop words etc. as they do not give us any sentiments.
5) Find out the opinion of users with the ratio of total number of tweets for an event to total number of tweets for all the events.
6) Determining the sentiment of each tweet through the AFFIN file.
7) Separate out tweets with polarity.
8) The results are determined when sentiment score are calculated considering the frequency of appearance of negative sentiments and positive sentiments.
9) Sentiments of each of the five events is generated for all the five countries.
10) Display the results in the form of bar-charts and pie-charts.

In this report, we show the demographic comparison between the opinions of users of five different countries on five different events. For instance, we want to understand and demonstrate whether the opinions formed about an event happening in one country by the people of different countries demographic based or not.

CHAPTER 4

IMPLEMENTATION DETAILS

In this chapter, the complete implementation process of our combined approach of sentiment and demographic analysis is define. We will show how to find out the latitude and longitude of a country. What is the process of collecting data. Preprocessing steps, calculating opinion percentage, determining sentiment score all will be discuss in this chapter.

Some other information like in which format our data is stored, language that we will use for implementation is also describe.

Following flow charts elucidate the steps involved in proposed methodology in order to properly understand the proposed work. It is shown in the below Figure 4.1.

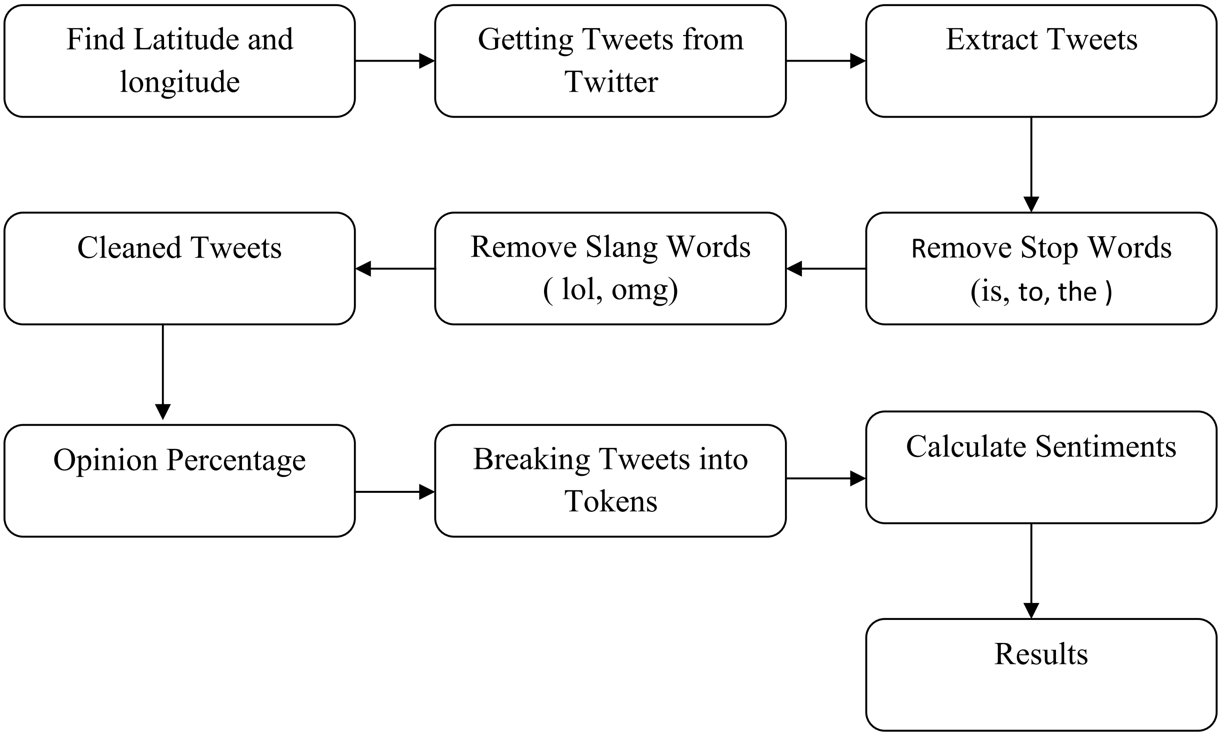
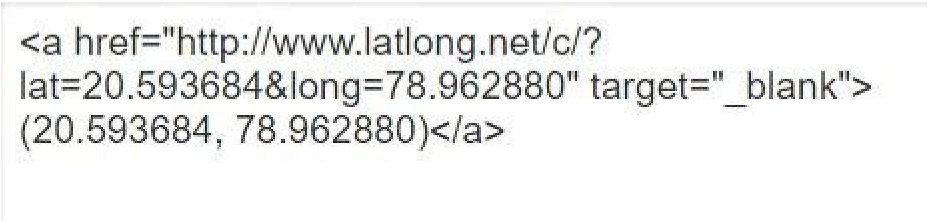

Find Latitude and longitude
Getting Tweets from Twitter
Extract Tweets
Remove Stop Words (is, to, the)
Remove Slang Words (lol, omg)
Cleaned Tweets
Opinion Percentage
Breaking Tweets into Tokens
Calculate Sentiments
Results

Fig. 4.1 Flow Chart of Complete Implementation Details

4.1 Latitude and Longitude

Latitude is an angle which ranges from 0° at the Equator to 90° at the poles. Lines of constant latitude, or parallels, run east–west as circles parallel to the equator. Latitude is used together with longitude to specify the precise location of features on the surface of the Earth. As of for the starting, we find out the latitude and longitude of a country. It is done through a service latlong.net [21]. It is an web services which helps in finding **lat long** of a required place, and receives its subsequent positions on map .Towns name or special places names are mostly used for searching purposes.



```
<a href="http://www.latlong.net/c/?
lat=20.593684&long=78.962880" target="_blank">
(20.593684, 78.962880)</a>
```

Fig. 4.2 Latitude/Longitude of India

4.2 Download Data

For performing Sentiment Analysis, Twitter data consisting of tweets are required for a particular event. For collecting the data and tweets we have used twitter public API. An Application Programming Interface(API) is a standardized system of instruction which help to get data from one platform or architecture to another platform and architecture, so that it can be used further. Our data contain 3,51,505 tweets for five different countries from the duration February 2016 to April 2016.

4.2.1 Web Scarping

Python 2.7 is installed to code the Twitter data mining, storage, retrieval and analysis. Tweepy is installed. It helps Python to talk with Twitter platform so that it can use its API.

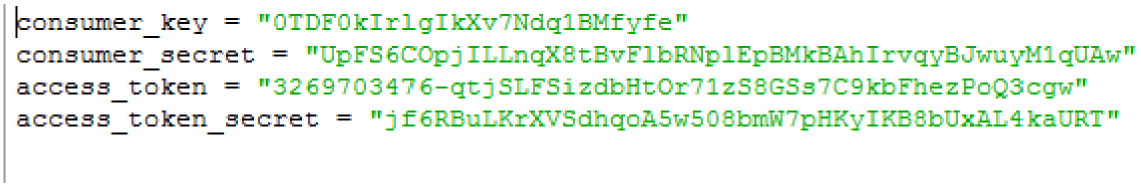
git clone http://github.com/tweepy/tweepy.git

python setup.py install

(or) pip install tweepy

4.2.2 Setup Twitter API Credential

Applications are created; “API Credentials” are setup in the Twitter Developers site so that twitter feed for sentiment analysis can be used [22]. Below are the following Twitter API Credentials for this report.

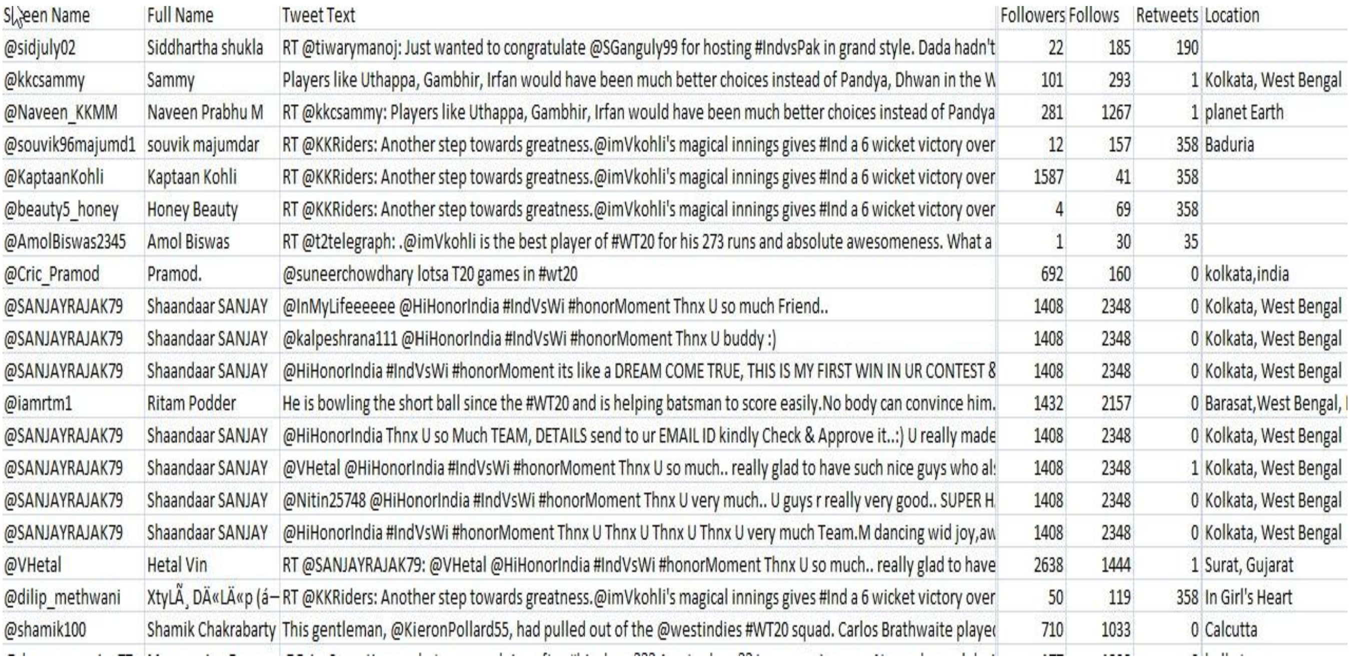


```
consumer_key = "0TDF0kIrlgIkXv7Ndq1BMfyfe"
consumer_secret = "UpFS6COpjILLnqX8tBvFlbRNplEpBMkBAhIrvqyBJwuyM1qUAw"
access_token = "3269703476-qtjSLFSizdbHtOr71zS8GSs7C9kbFhezPoQ3cgw"
access_token_secret = "jf6RBuLKrXVSdhqoA5w508bmW7pHKyIKB8bUxAL4kaURT"
```

Fig.4.3 Twitter API Credentials

Twitter has released its API for researchers and for web developers use. We have utilized the instruction mentioned within the Twitter API to crawl, collect and store information about users and tweets. To use twitter feed for sentiment analysis "API Credentials" are set up in the Twitter developer site by creating an application. For this purpose they provide us four keys: consumer_key, consumer_secret_key, access_ token_key, access_token_secret_key. After setting up the Twitter API Credentials, and retrieving latitude/longitude of a location from the latong.net online services, we start downloading the tweets.

With the help of Latitude and longitude we start downloading the tweets for each event. Data is downloaded in the form of comma separated values (csv) file. Initially data is in raw form. Lots of unwanted things are there in the data that we need to preprocess. It also contains other information with it like screen mane, full name, tweets, followers, location etc. that are shown in figure 4.4.

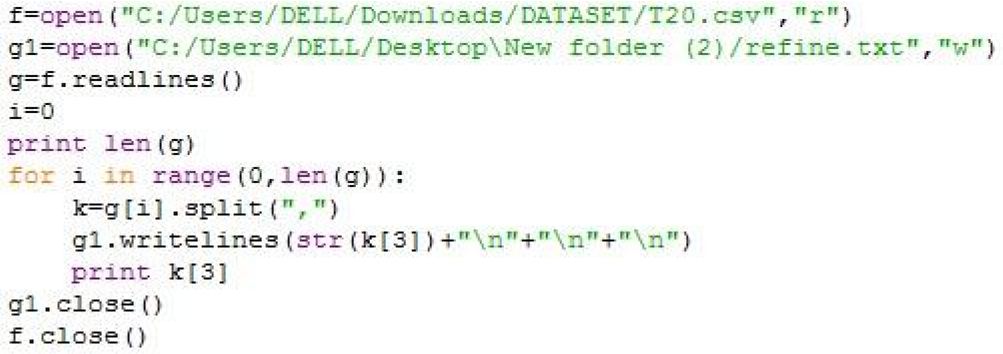


Screen Name	Full Name	Tweet Text	Followers	Follows	Retweets	Location
@sidjuly02	Siddhartha shukla	RT @tiwarymanoj: Just wanted to congratulate @SGanguly99 for hosting #IndvsPak in grand style. Dada hadn't	22	185	190	
@kkcsammy	Sammy	Players like Uthappa, Gambhir, Irfan would have been much better choices instead of Pandya, Dhwan in the W	101	293	1	Kolkata, West Bengal
@Naveen_KKMM	Naveen Prabhu M	RT @kkcsammy: Players like Uthappa, Gambhir, Irfan would have been much better choices instead of Pandya	281	1267	1	planet Earth
@souvik96majumd1	souvik majumdar	RT @KKRiders: Another step towards greatness.@imVkohli's magical innings gives #Ind a 6 wicket victory over	12	157	358	Baduria
@KaptaanKohli	Kaptaan Kohli	RT @KKRiders: Another step towards greatness.@imVkohli's magical innings gives #Ind a 6 wicket victory over	1587	41	358	
@beauty5_honey	Honey Beauty	RT @KKRiders: Another step towards greatness.@imVkohli's magical innings gives #Ind a 6 wicket victory over	4	69	358	
@AmolBiswas2345	Amol Biswas	RT @t2telegraph: .@imVkohli is the best player of #WT20 for his 273 runs and absolute awesomeness. What a	1	30	35	
@Cric_Pramod	Pramod.	@suneerchowdhary lotsa T20 games in #wt20	692	160	0	kolkata,india
@SANJAYRAJAK79	Shaandaar SANJAY	@InMyLifeeeeee @HiHonorIndia #IndVsWi #honorMoment Thnx U so much Friend..	1408	2348	0	Kolkata, West Bengal
@SANJAYRAJAK79	Shaandaar SANJAY	@kalpeshrana111 @HiHonorIndia #IndVsWi #honorMoment Thnx U buddy :)	1408	2348	0	Kolkata, West Bengal
@SANJAYRAJAK79	Shaandaar SANJAY	@HiHonorIndia #IndVsWi #honorMoment its like a DREAM COME TRUE, THIS IS MY FIRST WIN IN UR CONTEST 8	1408	2348	0	Kolkata, West Bengal
@iamrtm1	Ritam Podder	He is bowling the short ball since the #WT20 and is helping batsman to score easily.No body can convince him.	1432	2157	0	Barasat,West Bengal, I
@SANJAYRAJAK79	Shaandaar SANJAY	@HiHonorIndia Thnx U so Much TEAM, DETAILS send to ur EMAIL ID kindly Check & Approve it..:) U really made	1408	2348	0	Kolkata, West Bengal
@SANJAYRAJAK79	Shaandaar SANJAY	@VHetal @HiHonorIndia #IndVsWi #honorMoment Thnx U so much.. really glad to have such nice guys who al:	1408	2348	1	Kolkata, West Bengal
@SANJAYRAJAK79	Shaandaar SANJAY	@Nitin25748 @HiHonorIndia #IndVsWi #honorMoment Thnx U very much.. U guys r really very good.. SUPER H.	1408	2348	0	Kolkata, West Bengal
@SANJAYRAJAK79	Shaandaar SANJAY	@HiHonorIndia #IndVsWi #honorMoment Thnx U Thnx U Thnx U Thnx U very much Team.M dancing wid joy,aw	1408	2348	0	Kolkata, West Bengal
@VHetal	Hetal Vin	RT @SANJAYRAJAK79: @VHetal @HiHonorIndia #IndVsWi #honorMoment Thnx U so much.. really glad to have	2638	1444	1	Surat, Gujarat
@dilip_methwani	XtyLÃ̃¸ DÄ«LÄ«p (á–	RT @KKRiders: Another step towards greatness.@imVkohli's magical innings gives #Ind a 6 wicket victory over	50	119	358	In Girl's Heart
@shamik100	Shamik Chakrabarty	This gentleman, @KieronPollard55, had pulled out of the @westindies #WT20 squad. Carlos Brathwaite playec	710	1033	0	Calcutta

Fig 4.4: Unprocessed Twitter Feeds

4.3 Extract Tweets

After setting up the Twitter API credentials and downloading tweets related to a particular keyword for an event for which latitude and longitude information are known. The downloaded data contain the information like screen name, full name, tweet text, followers, follows, retweets and location. But for the work which has been proposed here, tweets and locations are the only requirement. Data has been refined and extracts the tweets from raw data.



```
f=open("C:/Users/DELL/Downloads/DATASET/T20.csv","r")
g1=open("C:/Users/DELL/Desktop\New folder (2)/refine.txt","w")
g=f.readlines()
i=0
print len(g)
for i in range(0,len(g)):
    k=g[i].split(",")
    g1.writelines(str(k[3])+"\n"+"\n"+"\n")
    print k[3]
g1.close()
f.close()
```

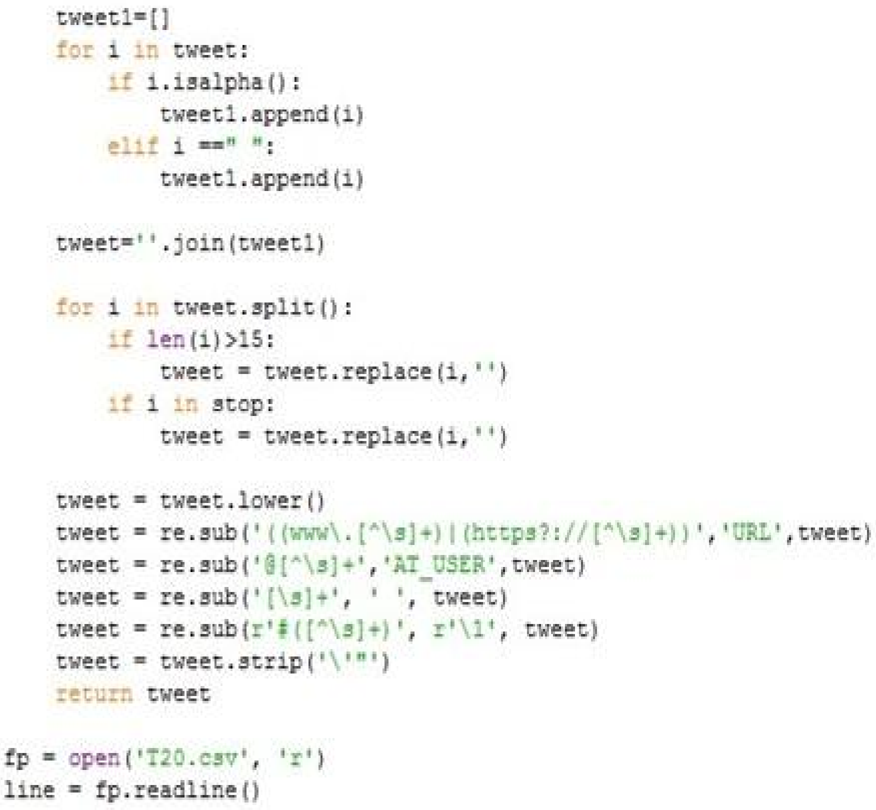
Fig 4.5 Refinement of Tweets

4.4 Preprocessing

In the previous sections we have described how Twitter data can be obtained using Twitter API and then procedure to extract tweets from the raw data. Extraction of Keywords is very tough in Twitter due to slang words which has been highly used these days and misspellings which is unfortunately another common features of most tweets. So a preprocessing step is required to avoid such errors prior to feature extraction. It is a process to remove the unwanted words from tweets that does not amounts to any sentiment. Before starting sentiment analysis, we need to do some data cleansing. We removed retweets (duplicates which do not add any value for our purpose) whose text starts with "RT". Some other preprocessing steps are described below:

- URLs does not signify any sentiment and replaced with word "URL".
- "#word" is replaced with "word".

- Slang words (e.g. lol, omg) are replaced with their actual phrase equivalences. A manually build slang dictionary is used for this purpose.
- Stop words (e.g. a, is, the) are removed since they does not indicate any sentiment.
- Replace repeated letters like huuuungry, huuungry, huuuuuuuuuuuungry into the token like huungry.
- Convert the tweets to lower case.
- Punctuations and additional whitespaces are removed. It is also helpful to replace multiple whitespaces with a single whitespace.



```
    tweet1=[]
    for i in tweet:
        if i.isalpha():
            tweet1.append(i)
        elif i ==" ":
            tweet1.append(i)

    tweet=''.join(tweet1)

    for i in tweet.split():
        if len(i)>15:
            tweet = tweet.replace(i,'')
        if i in stop:
            tweet = tweet.replace(i,'')

    tweet = tweet.lower()
    tweet = re.sub('((www\.[^\s]+)|(https?://[^\s]+))','URL',tweet)
    tweet = re.sub('@[^\s]+','AT_USER',tweet)
    tweet = re.sub('[\s]+', ' ', tweet)
    tweet = re.sub(r'#([^\s]+)', r'\1', tweet)
    tweet = tweet.strip('\'"')
    return tweet

fp = open('T20.csv', 'r')
line = fp.readline()
```

Fig4.6 Pseudo code for preprocessing steps

4.5 Opinion Mining

After completion the steps described in above sections, we find the opinion percentage of the each country users about each event. In this step we would find out the variations in the opinions of the users which is based on the fact that the particular event happens in their own country or on some other country. We have calculated the percentage of opinions by considering the total number of tweets of a country for an event and total number of tweets of all the countries for that particular event.

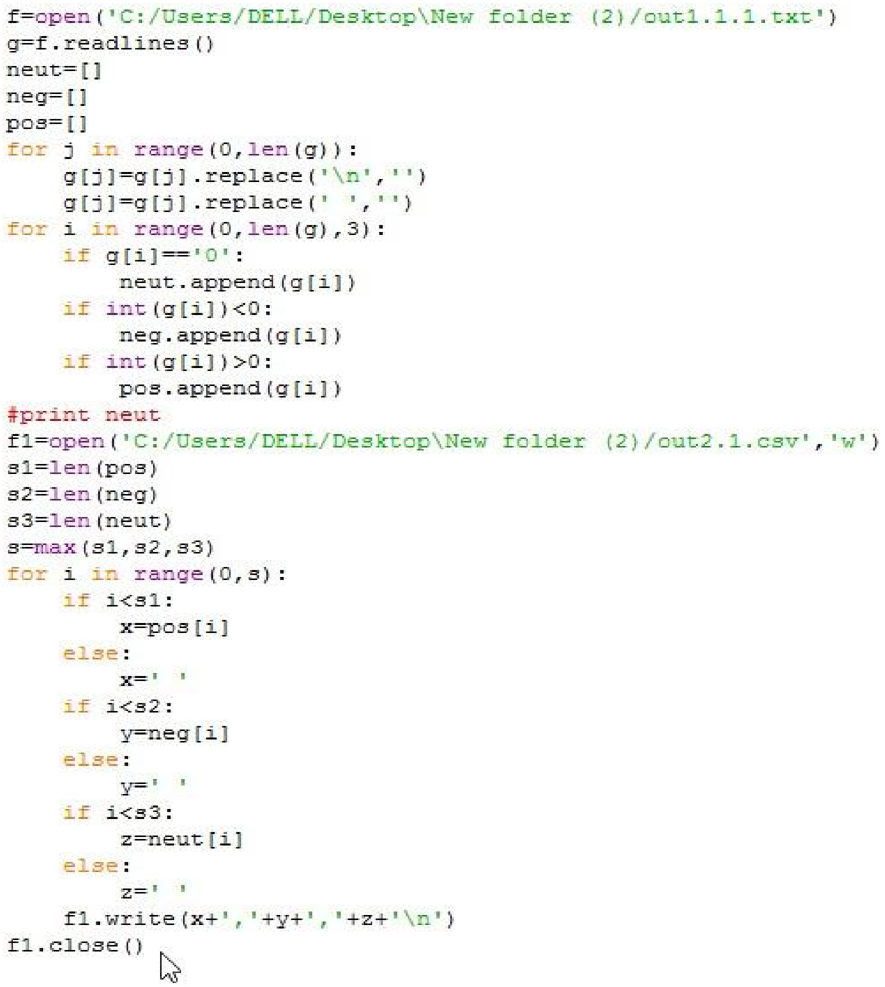
percentage of opinions = (total #tweets for a country / total # tweets for all countries) * 100

4.6 Determine the sentiments through AFFIN file

The general approach to perform sentiment analysis is based on the use of opinion words. It is used to express personal opinion about a certain event or product. Sentiment analysis is a process that calculates the sentiment score for tweets. Lexicon based approach is used for doing sentiment analysis. We have a text file which contains words with sentiment score.

AFINN is a list of English words rated for sentiment scores with an integer between -5 (negative) and +5 (positive). The words of AFINN file have been manually labeled by Finn Arup Nielson in 2009-2011. AFINN-111 version contains 2477 words and phrases. Applying the AFINN word list give a more graded response to textual sentiment analysis. Sentiments of posted tweets are calculated. It is largely based on sentiment score, which has been determined. Adding all tweets score to find its sum and this obtained sum is called sentiment tweet. Every Phrase or Words found in a tweet but not in AFINN-111 file has been assigned a score of 0. The AFINN-111.txt file format is tab-delimited. A tab character can be identified a "\t".

At first we tokenize the each tweet. With the help of AFFIN file each word get some sentiment score. After that by adding the score of all words of a tweet, we get sentiment score of complete tweet. Similar to this we get the sentiment score of a whole event by adding sentiment score of all the tweets of that event. Thus we get the positive and negative percentage of users' view about each event one by one.



```
f=open('C:/Users/DELL/Desktop\New folder (2)/out1.1.1.txt')
g=f.readlines()
neut=[]
neg=[]
pos=[]
for j in range(0,len(g)):
    g[j]=g[j].replace('\n','')
    g[j]=g[j].replace(' ','')
for i in range(0,len(g),3):
    if g[i]=='0':
        neut.append(g[i])
    if int(g[i])<0:
        neg.append(g[i])
    if int(g[i])>0:
        pos.append(g[i])
#print neut
f1=open('C:/Users/DELL/Desktop\New folder (2)/out2.1.csv','w')
s1=len(pos)
s2=len(neg)
s3=len(neut)
s=max(s1,s2,s3)
for i in range(0,s):
    if i<s1:
        x=pos[i]
    else:
        x=' '
    if i<s2:
        y=neg[i]
    else:
        y=' '
    if i<s3:
        z=neut[i]
    else:
        z=' '
    f1.write(x+','+y+','+z+'\n')
f1.close()
```

Fig 4.7 Calculate positive and negative polarity

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Results

Sentiment analysis has been carried out for five major events. It is done to assess users' broad opinion on social media. In addition, the variation of results based on the data collected is also discussed. Sentiment analysis for five different countries is done to know users' opinion for these countries different events. Using the user interface we have obtain the following results. Table 5.1 shows the percentage of overall opinions of users given by different countries on different events.

TABLE 5.1 Opinions of users in percentage

Events / Countries	Olympics	Oscar	T20	Paris Attack	Formula 1
US	21.65	48.30	17.35	36.09	38.16
India	27.14	14.39	38.96	17.28	12.93
France	07.28	12.29	14.96	18.09	11.53
Brazil	26.77	12.15	14.01	13.49	11.11
Australia	17.13	12.80	14.07	15.03	26.18

We can see in the above table that about the Olympics which is held on Brazil, the highest opinions are given by Indian Users and second highest by Brazilians. Similarly, about the Oscar which is held on United State, the highest opinions are given by US Users and second highest by Indians. T20 which is held on India , the highest opinions are given by Indian Users and second highest by US people. Paris attacks were a series of coordinated terrorist attacks occurred in

Paris, the highest opinions are given by US Users and second highest by French people. Similarly, Formula 1 Championship which is held on Australia, the highest opinions are given by US Users and second highest by Australians. We can see this result in figure 5.1, where the bar charts whose x-axes represents the country name and y-axes represents the opinion in percentage.

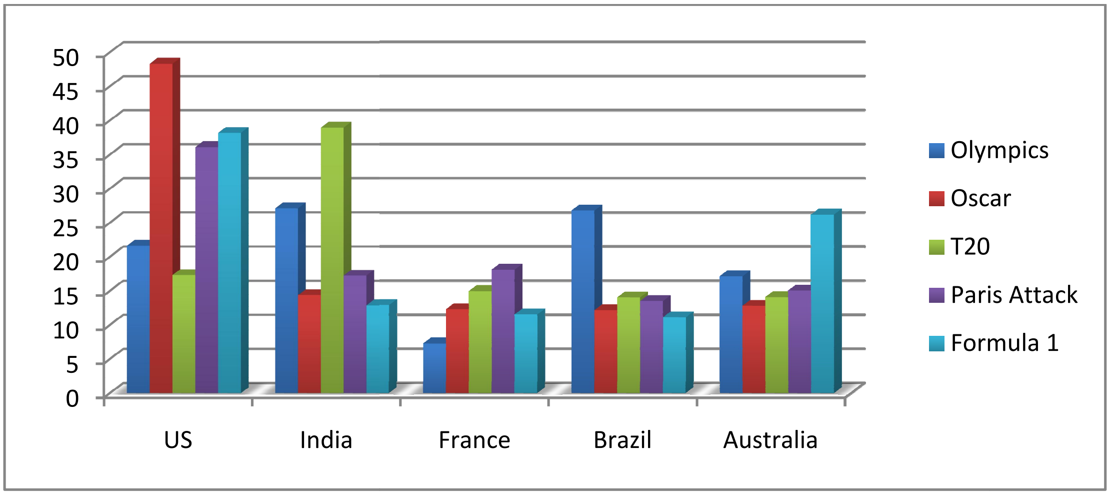
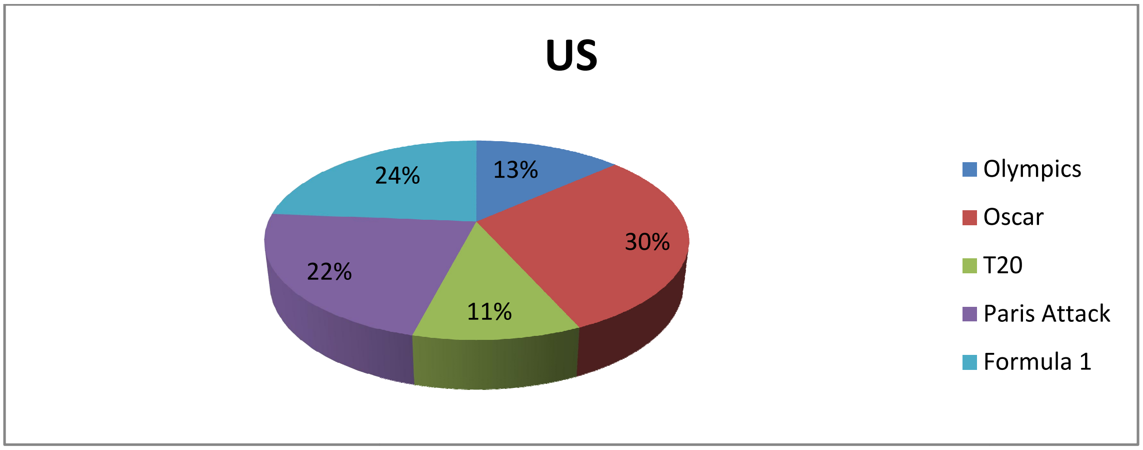
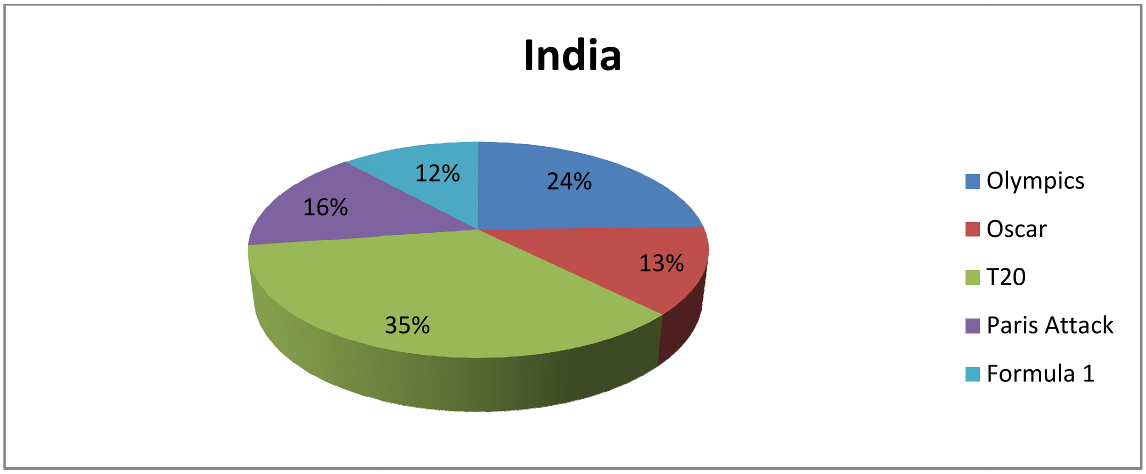


Fig. 5.1. Opinions of five countries about five events

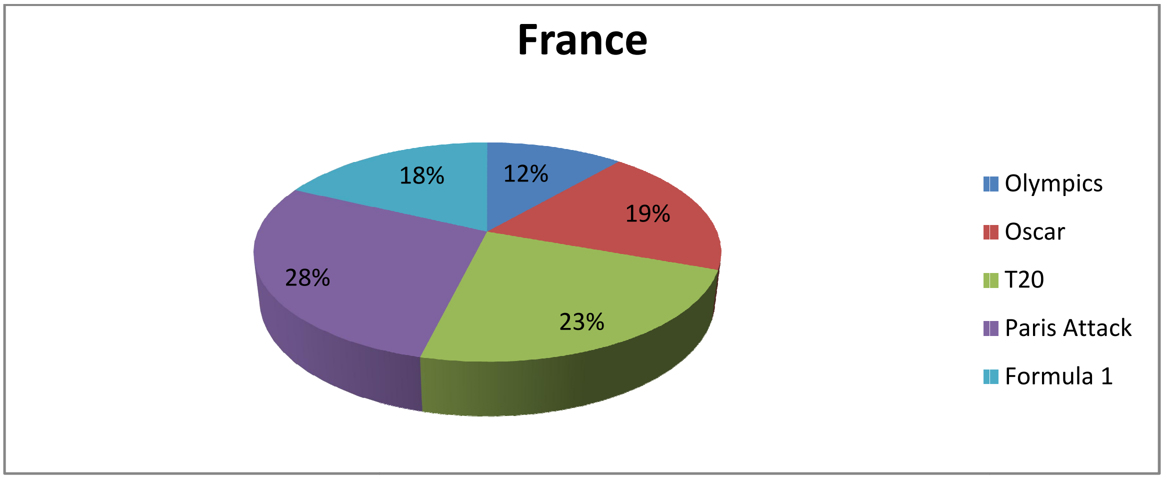
Figure 5.2 represents the percentage of each event given by a country. With the help of this we can see for which event a country give highest opinions. Like out of five different events in figure 5.2(a) US people gives their highest opinion about Oscar event. Similarly, India, France, Brazil and Australia's users share their more opinions about T20, Paris Attack, Olympic and Formula 1 Championship events.



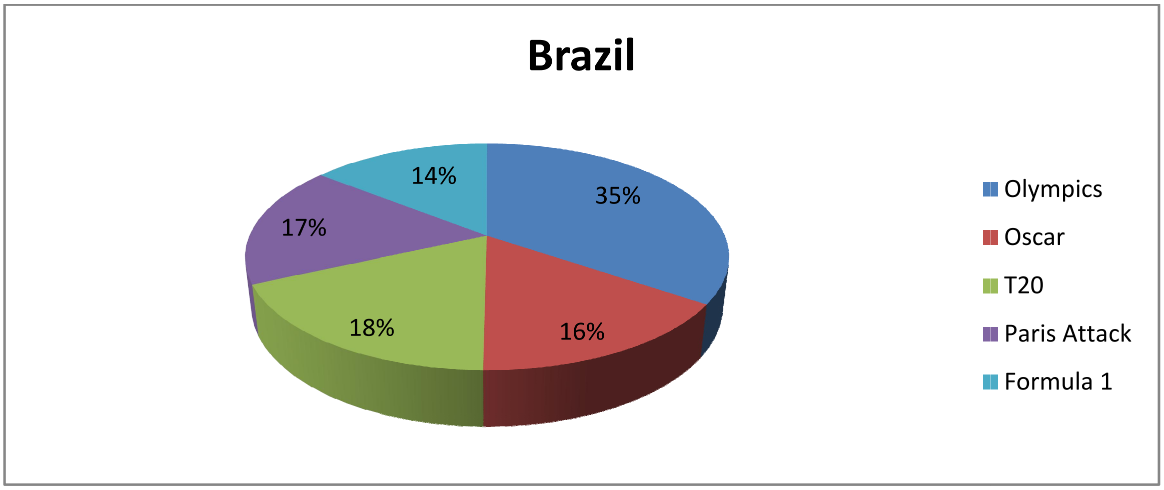
(a)


India
12%
24%
16%
13%
35%
Olympics
Oscar
T20
Paris Attack
Formula 1

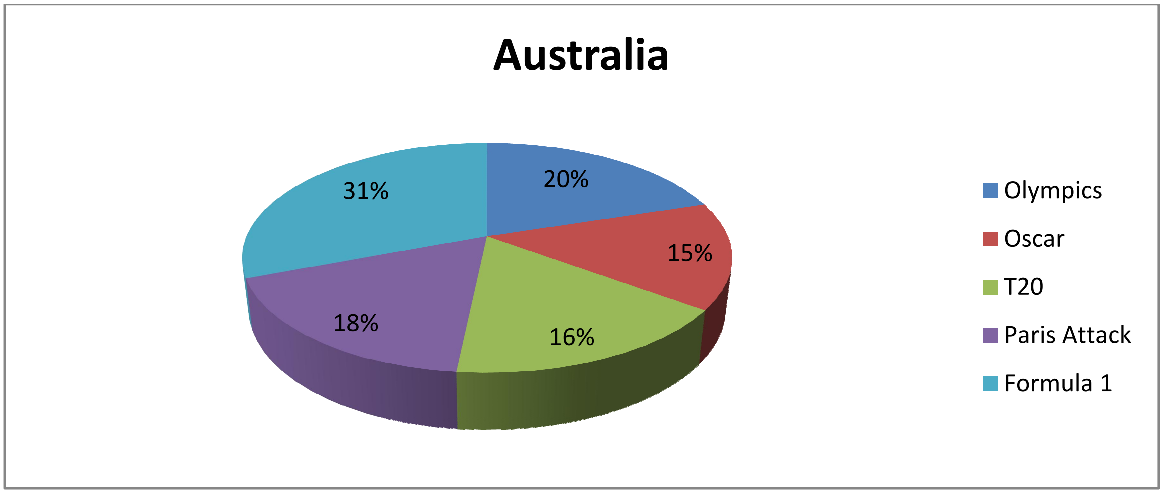
(b)


France
18%
12%
19%
28%
23%
Olympics
Oscar
T20
Paris Attack
Formula 1

(c)


Brazil
14%
35%
17%
18%
16%
Olympics
Oscar
T20
Paris Attack
Formula 1

(d)



(e)

Fig 5.2 (a-e) Percentage of each event given by each country

After getting opinions, the next we did is to find out the positive or negative views of users. With the help of Lexicon Based Approach we determine the positive and negative sentiments of people about different events. Table 5.2 shows Positive Sentiments of users in percentage. And table 5.3 shows the negative sentiments of users in percentage.

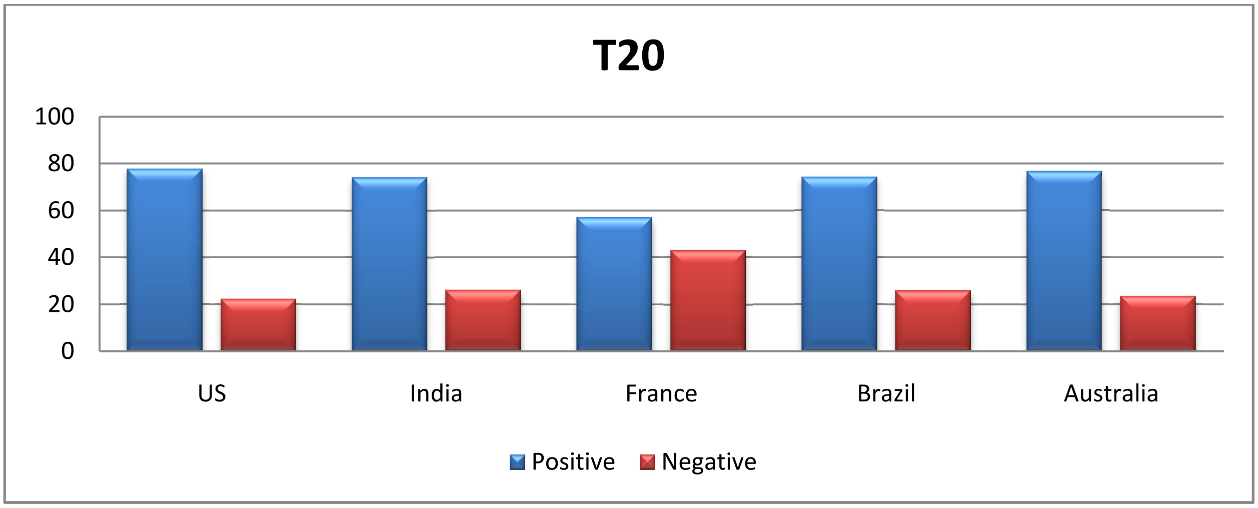
Table 5.2 Percentage of positive sentiments

Countries / Events	US	India	France	Brazil	Australia
T20	77.81	74.00	57.00	74.24	76.64
Oscar	75.96	82.47	97.61	81.73	83.63
Olympics	62.91	83.87	27.81	70.14	81.86
Paris Attack	40.94	28.18	48.17	68.46	74.60
Formula 1	60.99	82.00	58.55	67.90	84.00

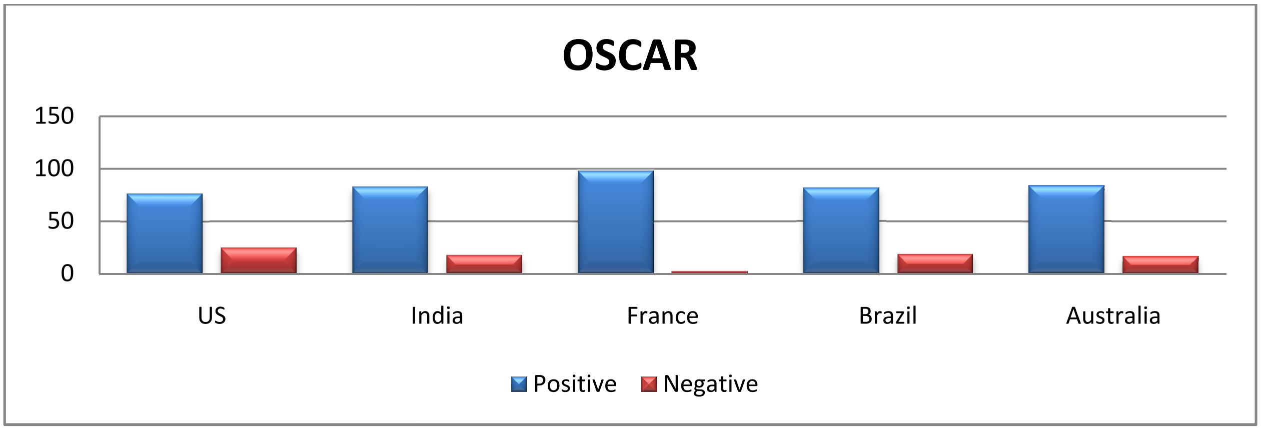
Table 5.3 Percentage of negative sentiments

Countries Events	US	India	France	Brazil	Australia
T20	22.18	25.99	42.99	25.7	23.35
Oscar	24.23	17.52	2.38	18.26	16.36
Olympics	37.08	16.12	72.18	29.85	18.13
Paris Attack	59.05	71.81	51.82	31.53	25.39
Formula 1	39.00	18.00	41.44	32.09	16.00

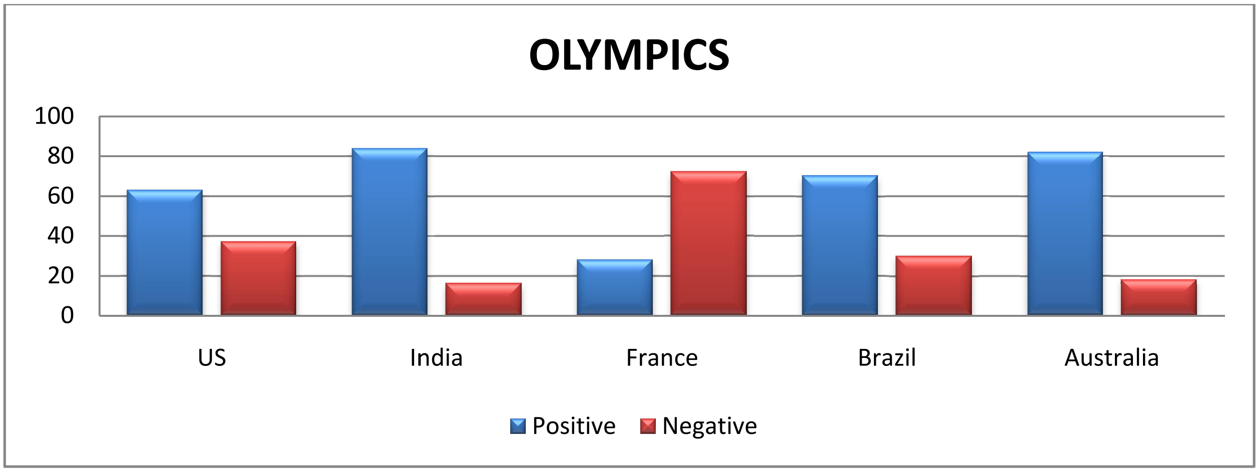
In figure 5.3(a-e) we shows the Sentiment Score of users in the form of graphs where x-axis shows the country names and y-axis shows the sentiment score in percentage.



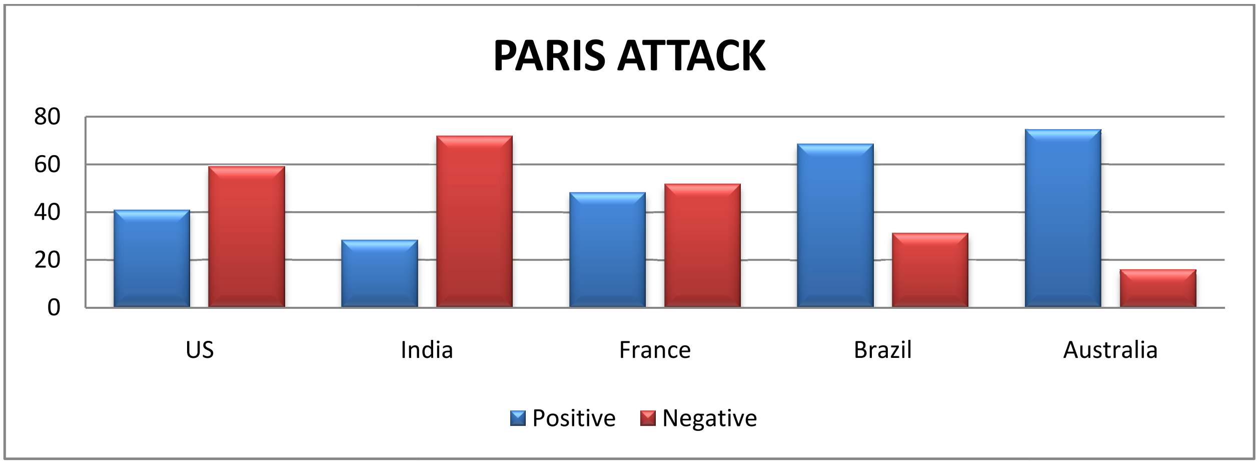
(a)


OSCAR
150
100
50
0
US
India
France
Brazil
Australia
🟦 Positive
🟥 Negative

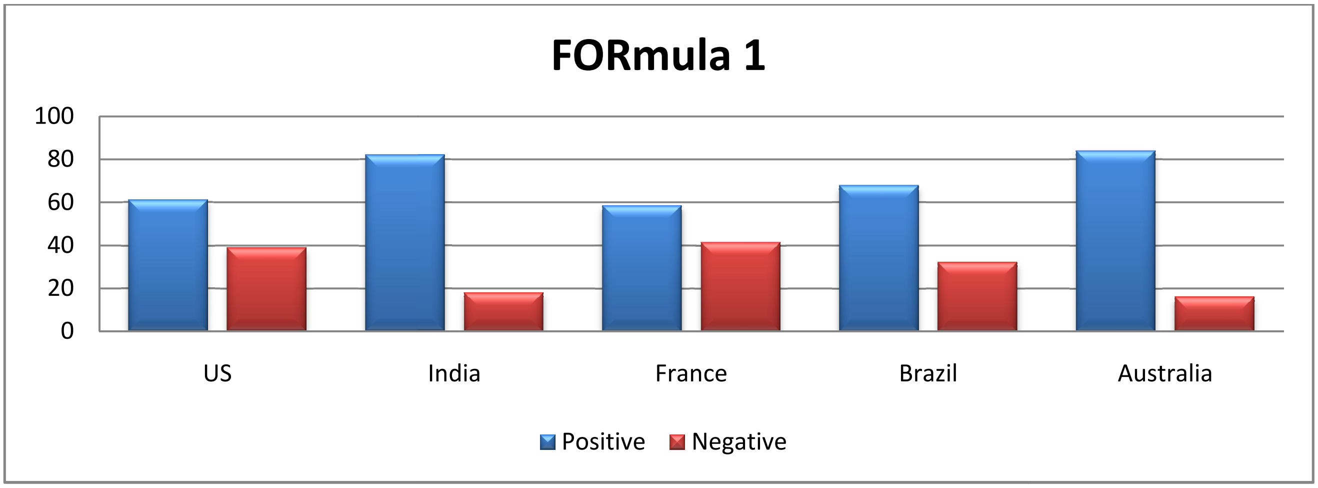
(b)


OLYMPICS
100
80
60
40
20
0
US
India
France
Brazil
Australia
🟦 Positive
🟥 Negative

(c)


PARIS ATTACK
80
60
40
20
0
US
India
France
Brazil
Australia
🟦 Positive
🟥 Negative

(d)



(e)

Fig 5.3(a-e) Positive and Negative sentiment score of users

5.2 Discussion

With the help of above results we can say that there are many factors that affect the results. Total population and Number of active users on Twitter of a country affect the results. Total number of tweets available and opinions are affected by the geolocations with users'. As users from many countries having higher number of tweets for a certain event. And for some it is very low or negligible tweets for the other events. Determining sentiments scores based on very less number of tweets can also hampers the results which can conclude to misleading information about the events happens at a particular location.

Table 5.1 above shows that no matter in which country an event is held, US users are in the highest or second highest position to give the opinions, because US is the number one country which is most active on twitter. 17% of US users are on twitter which is the highest across the world.

By looking on the results we can say that most of the time opinions of users are demographically based. If an event happens in a country then there are more chances that people of that country will tweet more about that event in place of other country events. But the sentiments of users across the world are almost same. For example, for Olympics, Oscar, T20 and Formula 1 championship there are more positive events. And for Paris Attack, which was an inhuman thing, more negative sentiments are given by people. We the help of this we can analyze the human behavior.

5.3 Limitations

There are many limitation in determining the patterns in twitter data, one such is that at any given instant only seven days of data can be collected hence it is required to collect data regularly and on continuous basis so that analysis over longer periods of time can be carried out. Which can led to a proper and realistic conclusion.

It may be possible that the current location from where a user tweet or the location that is mentioned in his twitter account is not the same. With this there is chances that the data that we collect is not sufficient to get appropriate results.

Language also plays critical role in this. Different countries have different languages, so at the time of data collection there is a possibility that we are not able to get sufficient amount of data from some countries because of language constraint.

CHAPTER 6

CONCLUSION AND FUTURE WORK

The demographic comparison on Twitter users has been done. In this report, we have done opinion mining and sentiment analysis on geotagged data. We have taken millions of tweets and with the help of these we can say that this research will help to know more about users' behavior.

Moreover, tweets occur in various geographic regions across the world can also helps in determination of the popularity of events considering those regions. A detailed analysis of tweets in order to determine statistically important user opinion requires a addressal of many parameter and criteria. These includes : a continuously large period of time for which tweets are aggregated to ensure representativeness, tweets making sufficient amount, best present geographic locations and an determination of some pattern or possible bias if the tweets comes largely from a particular mentioned geographic location.

In future, we can improve our work by finding the current and real location of user, which sometime creates misleading conclusion that the location from where a person tweet is that his real location or not. Other than this we can apply classification based on IP address. We hope this study enables further research in this area.

REFERENCES

[1] D. Arora, K. F. Li, and W. Neville, “Consumers' sentiment analysis of popular phone brands and operating systems preference using Twitter data: A feasibility study,” 2015 IEEE 29th International Conference on Advanced Information Networking and Applications (AINA), pp. 680-686, March 2015.

[2] A. Mislove, S. Lehmann, Y. Y. Ahn, J. P. Onnela and J. N. Rosenquist, “Understanding the Demographics of Twitter Users,” Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. AAAI Press, 2011. pp. 554-557, July 2011.

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, “Sentiment Analysis of Twitter Data,” LSM '11 Proceedings of the Workshop on Languages in Social Media, pp. 30-38, June 2011.

[4] L. Sloan, and J. Morgan, “Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter,” Published online 2015 Nov 6. doi: 10.1371/journal.pone.0142209.

[5] D. Murthy, A. Gross, and A. Pensavalle, “Urban Social Media Demographics: An Exploration of Twitter Use in Major American Cities,” vol. 21, pp. 33-49, November 2015.

[6] Neethu M. S., and Rajasree R., “Sentiment Analysis in Twitter using Machine Learning Techniques,” Computing, Communications and Networking Technologies (ICCCNT),2013 Fourth International Conference, 2013 IEEE, pp. 1-5, July 2013.

[7] A. Sarlan, C. Nadam, and S. Basri, “Twitter Sentiment Analysis,” Information Technology and Multimedia (ICIMU), 2014 International Conference, 2014 IEEE, pp. 212-216, November 2014.

[8] R. Srivastava, H. Kumar, M. P. S. Bhatia, and S. Jain, “Analyzing Delhi Assembly Election 2015 Using Textual Content of Social Network,” Proceedings of the Sixth International Conference on Computer and Communication Technology 2015, pp.78-85, September 2015.

[9] L. Sloan, J. Morgan, P. Bumap, and M. Williams, “Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data,” March 2015.

[10] S. Shaheidari, H. Dong, and Md Nor R. B. Daud, “Twitter Sentiment Mining: A multi domain analysis,” 2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, IEEE, pp.144-149, July 2013.

[11] T. McCormick, H. Lee, N. Cesare, and A. Shojaie, “Using Twitter for Demographic and Social Science Research: Tool for Data Collection,” Presented in Session 133: Social Media, Digital Tracks and Demography, October 2015.

[12] H. Oktay, A. Firat, and Z. Erterm, “Demograhic Breakdown of Twitter Users: An analysis based on names,” 2014 ASE BIGDATA/SOCIALCOM/CYBER SECURITY Conference, Stanford University, May 27-31, 2014.

[13] L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds and C. M. Danforth, “The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demography, and Objective Characteristics of Place,” PLoS ONE 8(5): e64417. doi:10.1371/journal.pone.0064417, May 2013.

[14] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, “Combining lexicon-based and learning-based methods for twitter sentiment analysis,” Hewlett-Packard Development Company, L.P, June 2011.

[15] M. Rambocas, and J. Gama, “Marketing Research: The Role of Sentiment Analysis,” The 5th SNA-KDD Workshop'11, University of Porto, 2013.

[16] D. Osimo, and F. Mureddu, “Research Challenges on Opinion Mining and Sentiment Analysis,” Proceeding of the 12th conference of Fruct association, 2010, United Kingdom.

[17] Z. Cheng, J. Caverlee, and K. Lee, “You are where you tweet: a content based approach to geo- locating twitter users,” CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 759-768, 2010.

[18] S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, and X. He, "Interpreting the Public Sentiment Variations on Twitter," IEEE Transactions on Knowledge and Data Engineering, Vol.26, pp. 1158-1170, July 2013.

[19] A. Hassan, A. Abbasi, and D. Zeng, "Twitter Sentiment Analysis: A Bootstrap Ensemble Framework," IEEE Social Computing (SocialCom), 2013 International Conference, pp. 357-364, September 2013.

[20] B. Gao, B. Berendt, and Vanschoren, "Who is More Positive in Private ? Analyzing Sentiment Differences across Privacy Levels and Demographic Factors in Facebook Chats and Posts," Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 605-610, 2015.

[21] www.latlong.net

[22] Twitter, "Twitter Developers," https://dev.twitter.com

[23] https://about.twitter.com/company

[24]www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research