Demographic Analysis of Twitter Users

Geeta
Department of Computer Science and Engineering
Indian Institute of Technology Roorkee
India-247667
Email: gtasngh20@gmail.com

Rajdeep Niyogi
Department of Computer Science and Engineering
Indian Institute of Technology Roorkee
India-247667
Email: rajdpfec@iitr.ac.in

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 paper, 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. We have performed this analysis on millions of twitter users residing in different locations and have demonstrated the findings using graphs and pie charts.

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

I. Introduction

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 .

There are about 320 million monthly and 100 million daily active users of twitter in the world. 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 paper, 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 in question. 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 approaches used for determining sentiments can be mainly divided into three categories: machine learning techniques and lexicon based techniques and hybrid techniques [1].

The remainder of the paper is organized as follows: in Section II, Related Work is discussed; in Section III, the proposed methodology and data collected from twitter is described; in Section IV, implementation details is discussed; in Section V, Results are discussed; finally, we conclude in Section VI.

II. RELATED WORK

Lots of research has been done on Sentiment Analysis and Opinion Mining and several different methods are proposed for this purpose. A few other studies have examined the demographics of social network users. Here are some literature work that is related to my work :

978-1-5090-2029-4/16/$31.00 @2016 IEEE

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.

Murthy, Gross and Pensavalle [5] shows us the invegtigation 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 approches: Power Law Behavior Method and Inter Tweet Interval Method.

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.

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.

III. PROPOSED METHODOLOGY

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.

A. Sentiment Analysis

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[1]. 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[3] or a particular clothing brand.

B. Demographic Analysis

The purpose of demographic analysis is to analyze the way the population of a particular area behaves[5]. 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[2]. 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 users, the common characteristics of group of users of twitter etc.

C. 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 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. The demographic information of a population can be useful for social scientists, marketers, and policy makers.

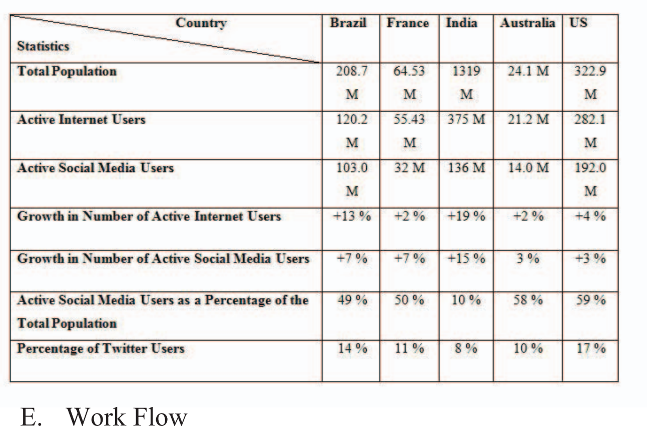
Most of the work done in the past is related to performing sentiment analysis in order to determine the sentiments of users about a product or brand or on finding demographic information of the users. However, these works do not study how users' opinion is affected by demographics. Or in other words 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.

D. Twitter Demographics

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. 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.

TABLE I. Demographic description of each of the five countries



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 %

E. Work Flow

The below figure 1 shows the basic flow diagram of our method.

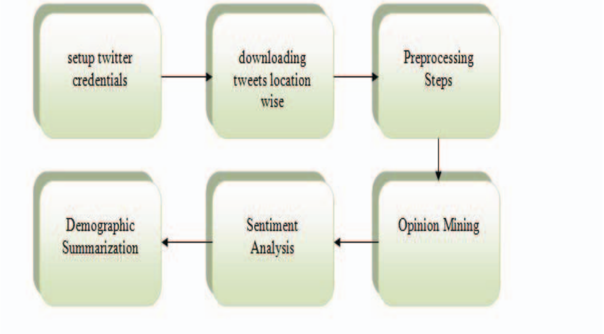

setup twitter credentials
downloading tweets location wise
Preprocessing Steps
Opinion Mining
Sentiment Analysis
Demographic Summarization

Fig. 1. Workflow overview of our approach

We have used Lexicon Based Sentiment Analysis Approach to express personal opinion about an event. In lexicon based analysis we look for opinion words in the data viz. tweets in order to determine whether a users' opinion is positive or negative. In this paper we show the demographic comparison as a result of five different countries with their five different events. For example we want to show that if an event is happen to a country then the opinion of other countries about that event is demographic based or not.

F. 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 shown with sentiment score calculated based on positive and negative 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.

IV. IMPLEMENTATION DETAILS

A. Latitude and Longitude

Twitter user 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 geo-tag his/her tweets. The geo-tagging feature in the Twitter API allows attaching location to a tweet. As of for the starting, we find out the latitude and longitude of a country. It is done through a service latlong.net[6]. 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.

B. 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 to another platform.

C. Web Scarping and setup Twitter API Credential

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. 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 setup 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.

D. Extract Tweets

After setting up the Twitter API credentials, we can download the 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.

E. Preprocessing

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.

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

F. 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 for a country and total number of tweets for all the countries.

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

G. Sentiments through AFINN

AFINN is a list of words that is written in English language for sentiment scores with values between -5 (negative) and +5 (positive). Words of this file have been manually. 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. In this step we separate out the positive and negative percentage of users' view about each event one by one.

V. 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 2 shows the percentage of overall opinions of users given by different countries on different events.

TABLE II. 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 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 Indians followed by US Users. Paris Attack 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 followed by Australians. In figure 2 the bar charts whose x-axes represents the country name and y-axes represents the opinion in percentage.

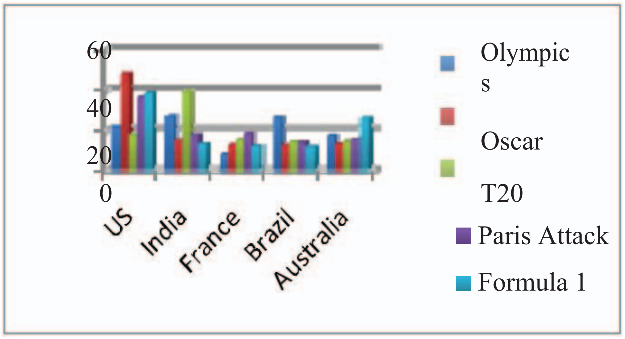


Fig. 2. Opinions of five countries about five events

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 3 and table 4 shows Positive and Negative Sentiments of users in percentage respectively.

TABLE III. Positive views of users for five different events

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 IV. Negative views of users for five different events

Countries / Events	US	India	France	Brazil	Australia
T20	22.18	25.99	42.99	25.75	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 figures 3-7 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.

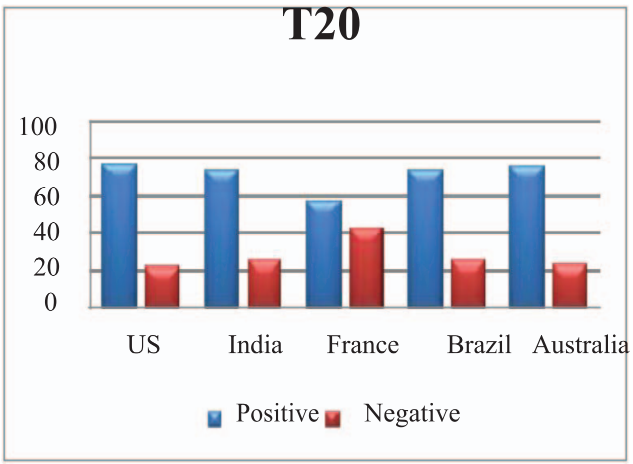


Fig. 3. Percentage of sentiment score about T20 for five countries

In figure 3, about the event T20 that held in India most of the tweets are positive not only in India but also in others countries too.

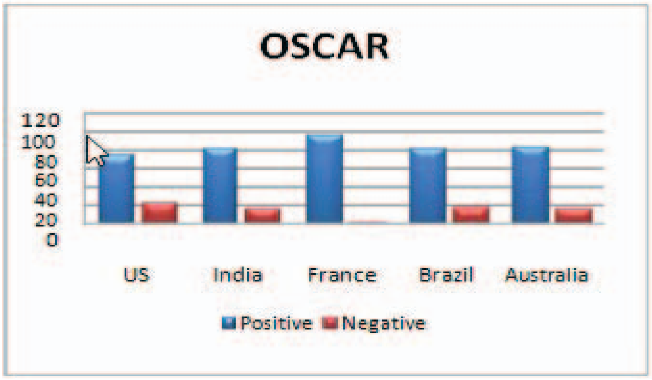


Fig. 4. Percentage of sentiment score about Oscar for five countries

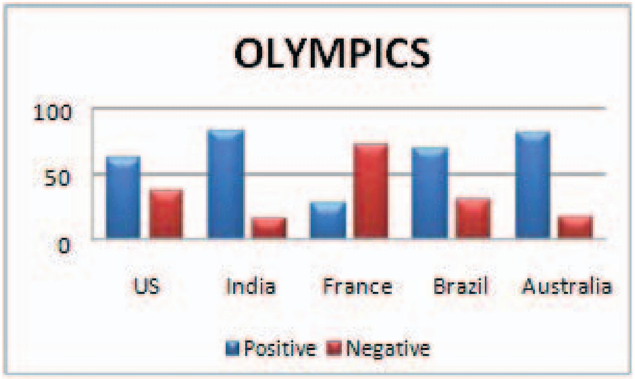


Fig. 5. Percentage of sentiment score about Olympics for five countries

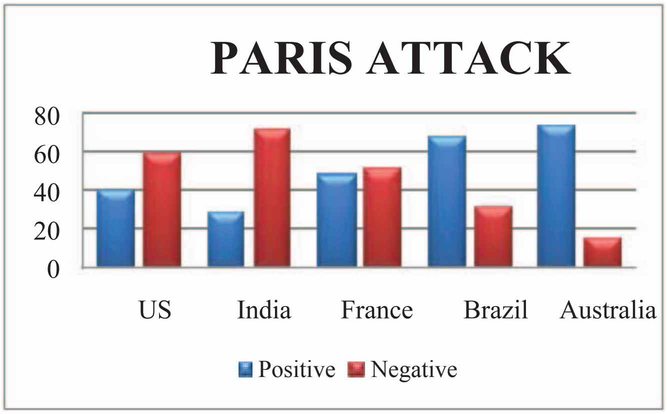


Fig. 6. Percentage of sentiment score about Paris Attack for five countries

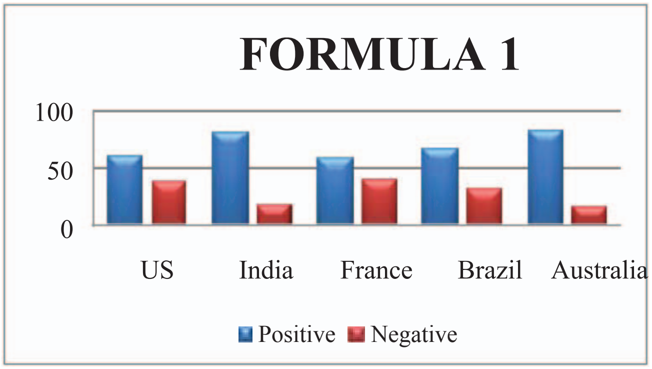


Fig. 7. Percentage of sentiment score about Formula 1 for five countries

With the help of above results we can say that there are many factors that affect the results. Total population and numbers of active users on Twitter of a country affect the results. Total number of tweets available and opinions are affected by the geo-locations. As users from many countries having higher number of tweets for a certain events. And for some it is very low and negligible tweets for the other events. Determining Sentiment score 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. 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's events. But the sentiments of users across the world are almost same. For example, for Olympics, Oscar, T20 and Formula 1 Championship we get more positive opinions. And for Paris Attack, which was an inhuman thing, more negative sentiments are given by people. With the help of this we can analyze the human behavior.

There are some limitation of this work which were identified based on the above results: 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. Second, 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

VI. Conclusion and Future Work

The demographic comparison on Twitter users has been done. In this paper, we have done opinion mining and sentiment analysis on geo-tagged data. Based on the results we have seen that most of the time users' opinion are demographic based but not in all cases. 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.

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, 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,” July 2011.

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, “Sentiment Analysis of Twitter Data,” pp. 30-38, June 2011.

[4] L. Sloan, and J. Morgan, “Who Tweets witj Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter,” November 2015.

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

[6] www.latlong.net