SYSTEMATIC REVIEW OF LITERATURE USING TWITTER AS A TOOL

by

Mudit Pradyumn

July, 2018

Director of Thesis: Nasseh Tabrizi, PhD

Major Department: Computer Science

Twitter has over 330 million active monthly users producing roughly 500 million Tweets per day, or 200 billion Tweets a year. Making this one of the largest human-generated opinion data collections. In addition to this major advantage, Twitter generates real-time data, making it possible to gain insights on trending information instantaneously. People post about a wide variety of subjects, including their opinions, feelings, situations, current trends, and products. This makes it a great data source for analyzing the sentiments of people on a variety of subjects. In this study, out of 1025 research papers on Twitter data analytics from 2011-2017, papers from only 20 selected journals were considered for review. They were then classified based on their year of publication, their titles, data mining methods, and application areas. In the course of this study a tool for the Sentiment Analysis of the Twitter data was developed and used to conduct a case study on individuals on marijuana use during pregnancy.

SYSTEMATIC REVIEW OF LITERATURE USING TWITTER AS A TOOL

A Thesis

Presented to The Faculty of the Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Master of Science in Software Engineering

by

Mudit Pradyumn

July, 2018

SYSTEMATIC REVIEW OF LITERATURE USING TWITTER AS A TOOL

by

Mudit Pradyumn

APPROVED BY:

DIRECTOR OF THESIS: _____

Nasseh Tabrizi, PhD

COMMITTEE MEMBER: _____

Venkat Gudivada, PhD

COMMITTEE MEMBER: _____

Mark Hills, PhD

CHAIR OF THE DEPARTMENT

OF COMPUTER SCIENCE: _____ Venkat Gudivada, PhD

DEAN OF THE

GRADUATE SCHOOL: _____ Paul J. Gemperline, PhD

Table of Contents

LIST OF TABLES vi

LIST OF FIGURES vii

1 INTRODUCTION 1

2 RELATED WORK 4

3 SYSTEMATIC REVIEW OF LITERATURE AND RESULTS 7

3.1 Overview 7

3.1.1 Sentiment Analysis 7

3.1.2 Linguistic Analysis 13

3.1.3 Comparative Analysis 15

3.1.4 Influencer Identification 16

3.1.5 Disaster Management 17

3.1.6 Cybercrime Detection 18

3.1.7 Public Health Service 19

3.1.8 Disease Management 20

3.1.9 Future Prediction 20

3.1.10 Medical Complaints 21

3.1.11 Computer Communications 22

3.2 Review of the Results 23

4 DEVELOPMENT OF SENTIMENT ANALYSIS FRAMEWORK AND CASE STUDY 27

4.1 Twitter Sentiment Analysis Framework 27

4.1.1 Description of Tools used in the Framework 27

Django 27

Amazon Web Services (AWS) 28

Tweepy 28

TextBlob 28

4.1.2 Verification of the Framework 29

4.1.3 Framework 30

4.1.4 Data Extraction 31

4.1.5 Sentiment Analysis 31

4.1.6 Data Visualization 32

4.2 Case Study 32

5 CONCLUSION AND FUTURE WORK 39

BIBLIOGRAPHY 42

LIST OF TABLES

3.1 Selection of papers based on field of research 24

3.2 Selection of papers based on year of publication 25

3.3 Selection of papers based on journal 26

4.1 Manually analyzed tweets 34

LIST OF FIGURES

3.1 Flow diagram showing screening of the research papers 23

4.1 Tokenization of words in sentences 29

4.2 Framework of the application 30

4.3 Visualization of data 32

4.4 Sentiments based on keywords 34

4.5 Positive sentiments 35

4.6 Negative sentiments 35

Chapter 1: Introduction

Presently, a tremendous amount of digital information, namely sensor data, social media data, data storage, public web among others are available at our fingertips. The rapid accumulation of such data requires specialized methods and techniques to extract meaningful information and detect significant and interesting trends to make sense of the bigger picture to be utilized effectively in finding solutions to everyday problems. As the technology is evolving, new and effective methods of Big Data analytics [1] are being developed.

Today we are generating data at much faster rate than we ever have in history. If we look around us, we see that there are various sources of data generation. Consider the example of smart phones which captures our day to day data in the form of text, photos, videos, our location and much more. Similarly, a lot of data is generated from activities like shopping, conversations, sensors, and Internet of Things (IoT) [2]. Acording to [3] we were creating more than 2.5 Billion GB of data every day, a jet engine alone produce up to half a terabyte of data during flight [4]. This number is expected to generate more than a 4.4 zettabytes by 2020. Many industries have adopted Big Data analytics and are now reaping the benefits of Big Data in different ways. For example, industries utilizing Big Data are likely to be three times faster and more accurate in determining patterns of growth and decision making. Further study [3] has revealed that by integrating Big Data analytics into healthcare, the industries could save up to $450 Billion, a savings of more than $1000 per person per

year.

Twitter is an example of a popular source of Big Data. This is one of the largest collections of human generated data, and it provides a lot of benefits to users. Twitter has 330 million active monthly users making about 500 million Tweets, accounting to 200 Billion Tweets per year. A potential benefit of using Twitter is that it generates real time data that is publicly available. According to the IRB the federal definition of a human subject is a living individual about whom an investigator (whether professional or student) conducting research obtains data through intervention or interaction with the individual, or identifiable private information. In order to confirm we consulted with the IRB office and found that data provided by Twitter is public and therefore does not fit in this criterion. This thesis begins with the introduction to the major fields in which Twitter data is used as a means of research. These researches are grouped on the bases of field of research such as Sentiment Analysis [5], Linguistic Analysis, Comparison of Data, Influencer Identification, Disaster Management Processing, Information System, Impact on People, Detecting Relevant Topic, Computer Communication, Cybercrime Detection, Public Health Service, Disease Management, Future Event Prediction and Medical Complaints.

In this study, research papers were grouped based on the year of publication, the journal in which these research papers were published, their field of research and relevance to the topic.

Our work was concluded with a case study where a Sentiment Analysis tool was developed to analyze the sentiments of people regarding the use of marijuana by pregnant women. Tweets related to the issue were collected as the use of marijuana has been on the rise over the past few years and is increasing as fast as the Internet was in the 2000s. Forbes suggests that the marijuana market increased more than 30% in the year of 2016 and is valued at 6.7 Billion US Dollars in the US and Canada. It is

predicted to reach up to 20.2 Billion dollars by the year 2021. The usage of marijuana and whether it has any side effects has been an issue for very long time. The fact that there are no conclusive reports yet on whether it has harmful effects or not and the strict regulations regarding its usage has created even more confusion in peoples minds. The situation becomes trickier when pregnant women raise questions about its usage during pregnancy. There are three major problems to note here, doctors are reluctant to say anything about it, it is hard to determine what people really feel about it and do during pregnancy, and due to stringent laws on doing experiments on pregnant women it is hard to determine the effects of it on the newborn. If we start conducting research on this issue by surveying people, it will take a long time from a large section of society and, the sentiments of people tend to change with time. The purpose of the case study reported in this thesis is to analyze peoples sentiments associated with the use of marijuana during pregnancy, to analyze the effectiveness of the methods generally used by researchers and to verify the developed tool for effective search results. Sentiment Analysis or opinion mining is a process of analyzing if the given sentence is positive, negative, or neutral. There are generally two types of Sentiment Analysis tools, based on the methods they use to analyze the sentence. One type of Sentiment Analysis tool analyzes the statement at various levels, the sentiment of the whole document is analyzed by the phrase, and the individual words used in the sentence. In the second type, the tool only focuses on individual words. This means if someone says, I love marijuana, but it is bad when you are pregnant; the tool will look at the sentiment related words love and hate from its collection. Love, being positive and hate being negative it will rate the sentence as neutral. Whereas if the whole phrase is analyzed, the importance of "but" in the sentence is realized and the tool gives two different sentiments for the statement.

Chapter 2: Related Work

The related work section covers the other aspects of Twitter data usage, with an entirely different approach as discussed in the thesis. An analysis of Big Data technologies InfoSphere BigInsights and Apache Flume [6] was conducted by Birjali et al. Multiple sets of data for various research purposes was first collected from Twitter by Apache Flume, stored in Hadoop, and then displayed with BigSheets after being analyzed using InfoSphere BigInsights. They chose Twitter as their Big Data source, due to the increasingly large amount of data generated daily by its users. This method uses the Hadoop Distributed File System (HDSF) in order to utilize the MapReduce feature, enabling the collection of larger data sets (Tweets). MapReduce counts the number of times a matching data set is iterated and then displays the results. Apaches Flume Next Generation (NG) was used to collect the Tweets used in this case study. Flume NG uses a process that first collects data (Tweets) from multiple sources and holds them in memory, and then stores them in the HDSF using JAQL script, which is a data processing and query language. After a thorough examination of InfoSphere BigInsights analytics, a separate data collection tool developed from Apache Flume was tested, and the results were analyzed using InfoSphere BigInsights. It was determined that the technique used by the tool developed from Apache Flume was not only superior to older methods, but faster as well.

A paper on the Intelligent Mining of Public Social Networks Influence in Society (MISNIS) tool [7] highlights several key limitations on current methods, such as Twit-

ter API restrictions and dependency on hashtags and keywords for categorization, and demonstrates how MISNIS overcomes these limitations, increasing productivity by 80% and 40% respectively. MISNIS uses polarity sentiment analysis, and does not use a language dependent lexicon. While this approach is limiting, it does not negate MISNISs apparent superiority, and is open to further development in future.

Joao P. Carvalho and his collaborators [7] demonstrate MISNIS by applying it to track, catalogue, analyze, and trace current events in Portugal; however, MISNIS can be applied in many other fields with various other research questions. It can collect, store, manage, mine, and display data by using Computational Intelligence, Information Retrieval, Big Data, Topic Detection, User Influence and Sentiment Analysis. This method uses geolocation to collect Tweets within Twitter's API restriction of 1% data collection, then traces the collected Tweets back to the users accounts to collect additional Tweets that meet the search criteria from multiple Twitter API accounts. A file of every viable user was created and maintained to facilitate this process. Mongo DB was used for all data storage, and a REST API was used to handle the data once it was collected. In addition, the REST API is also the tool used to collect data from individual users. This method does not make collection limitless, as it is also minimally restricted by Twitter.

An insightful exploratory analyzer, demonstrates the capabilities of Tweets Characterization Methodology (TCHARM) [8] to organize collected Tweets based on geographical location, the time of the Tweet, as well as its contents. TCHARM uses the Text And Spatio-TEmporal (TAST) distance measure in order to group similar Tweets based on all three categories. This means that TCHARM is capable of grouping Tweets about the same, or similar subjects, from geographically close, or specified regions, that were Tweeted around the same time. The case study conducted in this paper to demonstrate TCHARMSs performance searched for and categorized Tweets

related to the 2014 FIFA world cup. Through this study it was determined that the TAST feature utilized by TCHARM produced a more even distribution of the three factors tested for by TCHARM than did other methods. The authors also address avenues for future work based on TCHARMs limitations. One such limitation is the length of time it takes to set the specifications of TCHARMs features. It is also suggested that the K-means algorithm used by TCHARM may collect too broad a range of Tweets containing the three factors for categorization. While this means that some collections of Tweets are more loosely related than is desirable, it does not affect the overall higher efficiency demonstrated by this method. TCHARM can handle a high number of Tweets in its data collection due to its use of Apache Spark as its platform, and collects Tweets quickly on an hour to hour recurring basis.

Chapter 3: Systematic Review of Literature and Results

3.1 Overview

As the number of people associated with the social media is increasing we get to see millions of people generating data over social media. Thus, tons of data is available on social media from which a variety of useful information can be derived. Through analysis of this data, we can learn about their lives, opinion, likes, dislikes and what they see around them. A lot of research is going in this field and people have come up with some interesting results in different areas. Twitter is one such commonly used source where people share their status with the public. Twitter is also utilized to understand what people think about certain issues, determine their location, their language and dialects, their opinion, and the impact of different events on people. In this study we discuss the use of Twitter data in different fields [1] such as Sentiment Analysis, Language, Comparative Analysis, Influencer Identification, Disaster Management, Information Systems, Impact on People, cybercrime Detection, Public Health, Disease Management, Future Prediction, Medical Complaints and Computer Communication.

3.1.1 Sentiment Analysis

In this research Tweeted content was analyzed using open tool Sentiment Analysis. The research was conducted for a single country i.e. Turkey to check the gross national

happiness [9]. Twenty thousand people posted 35 million Tweets that were then analyzed by an open source Sentiment Analysis tool and compared with data from previous years. Based on level of happiness they were categorized as happy, neutral or negative. Stock index and the happiness of Twitter users were analyzed to see if there was any relationship between the two. It was revealed that there is a significant relationship between happiness and stock index. Thus, the sentiments of people can be determined by the nature of Tweets they post. Since the study belongs to one country, it may not be true in equal measures in other countries at a global level. In other research, if the attitude of a person contemplating suicide is detected, he can be counseled, and his behavior can be modified to prevent suicidal attempts in future. Based on the keywords like suicidal; suicide; kill myself; internecine; my suicide note; my suicide letter; sleep forever; better off; dead; suicide plan; suicide pact; tired of living, and die alone different levels for risk of suicide were determined by the researchers. The research showed that people do use Twitter data to express their suicidal feelings among other sentimental expressions [10]. However, using this model, we may not be able to determine the age and gender of the Twitter user.

Sports occupy an important position in many peoples lives. Various sentiments such as excitement, disappointment, thrill, joy, inspiration, fraternity, etc. are expressed in Tweets related to sports activities. World Cup 2014 in the Twitter world brought out a variety of public moods [11]. Hash tags like #FIFA, #Football, #World cup, and #Soccer were searched for using a web crawler and Twitters Search API. Data was collected during five games of the World Cup. Three of the games were U.S. games while two were randomly selected games for comparison. Python and the Natural Language Toolkit were the platforms used for the project. URLs, user names, and hash tags were removed from the Tweets for the sake of convenience. Emotions such as anger, joy, sadness, disgust, and surprise were measured by the

NRC Word-Emotion Association lexicon. R was used to write an application that could calculate the emotional score of a word. Emotions were also analyzed to gain further accuracy in measuring the emotions in a Tweet. As predicted, emotions of fear and anger peaked when events were not in favor of the U.S. soccer team. After Portugal scored a goal in one of the games, Tweets showing fear and anger rose greatly. Anticipation also increased after the U.S. gave up first goal. Anticipation decreased after the U.S. tied games and would increase again after the U.S. would score. The motional patterns displayed in the Tweet research were consistent with the researchers hypothesis. The system used in this study was designed to adapt its behavior based on the incoming Tweets. The Soft Frequent Pattern Mining (SFPM) algorithm has superior performance in detecting relevant topics when compared to other techniques. The SFPM allows users to easily change and update the way in which current terms are picked up while still being efficient. The research was conducted in real time and provides real-life reaction analysis. The expectations of the researchers were consistent with the Big Data analysis of the Tweets and the results can be logically explained.

However, the research was only conducted on English Tweets, which is extremely limiting due to the amount of foreign, non-English interest in the World Cup. Many other factors such as, commentary, shots on goal, foul calls, and missed shots can change someone's emotional reaction during the game.

This study set out to observe how public sentiment gathered through Twitter affects the performance of the stock market. The researchers were successfully able to create forecasting models that identified new variables and indicators of the movement of a certain stock within the stock market. They found that the sentiment of a Tweet is less useful in terms of prediction than the number of Tweets posted by a user [12]. Data Sift was used to collect only English Tweets between September 24 to November

24, 2014 from Facebook, Google, and Apple. In addition, the prices for the stocks were collected from Bloomberg. To filter out the meaningless noise on Twitter, only Tweets that included the companies ticker were observed. After collecting the Tweets, their sentiment was scored on a -20 to 20 scale based on the positivity or negativity of the Tweet. A klout score [13], a score that shows the level of influence an individual has, was also calculated for each Tweet. An ordinary least square regression and a linear probability model were used to review the relations between the stocks and the sentiments of the Tweets. Specific variables were present in all the regression models. The study was conducted over a short time frame with only a small amount of companies.

Tweets on a company's events can enhance its market scope and stock value [14]. Four text analysis tools were used to look for keywords. People usually express their thoughts through Twitter on an issue, and identifying these feelings in a correct manner through keywords is possible up to certain level. However, there are many challenges faced in this field due to the irrelevant Tweets containing similar keywords or due to sarcasm in the language. The analysis of the Tweets on a company's events and peoples response in the financial market narrows down the spectrum of the research to just the community related to finance. The investors want to gain maximum profit out of their investment, but there is no certain way by which it can be assured. Conventionally, the technique used to analyze the future trend of the market is based on the previously available data and the similar trend followed in those situations. The requirement of a product was to analyze the market, and conclusions were derived from those analysis. This trend is currently changing as we are advancing into the internet age. We can now see the direct impact of events with the help of Tweets. As there is lot of clutter in Twitter data, the analysis must be in a very specific field, which can help bring in better and more relevant analysis. This

is done by selecting a company to be analyzed. A system is setup to extract Tweets related to the companys financial related aspects for analysis. Then a sentiment analyzing process is developed to extract meaningful information. This creates a picture showing where the company is leading or lagging. The main advantage of this method is the massive amount of data analyzed to give quality results.

Analysis of Tweet content verbatim misses the connotations of evolving Tweet language. The Sentiment Analysis remains one of the most challenging fields of research for predictions due to its need to find near accurate results. Various methods and strategies have been suggested for the analysis of sentiments attached to the Tweets. One of these methods is accomplished by creating a model which analyzes the Tweet based on the dictionary meaning of the words used, machine learning based principles and a mixed approach of these two [15]. Also, there are commonly accepted and used symbols or the representation of expression, such as use of a colon and a closing bracket to represent a smile of happiness. This helps in understanding the motive. Another challenge in understanding the motive behind Tweet arises when people use sarcastic or ironic language. When expressing certain thoughts, the user may use words which have an entirely different intent than apparent. To overcome this challenge the use of the commonly expressed symbols, the emoticons used with the Tweet, are analyzed. This method has proved successful to a certain degree. Now the trend of Sentiment Analysis is shifting from traditional Twitter analysis to creating quality data sets based on their accuracy. This is determined by multiple evaluations and trials at various levels and comparing them with the actual Tweets to see how they stand. Cuckoo search and two n-grams method were used to solve the problem to some extent. This has proven to be very effective in determining a conclusive review of Twitter data, but it is deficient in understanding sarcasm and irony.

With the increase in use of the Internet and social media, the micro-blog data and blog sentiment provide a useful material which can be used for stock market prediction, its volatility and survey sentiments [16]. The Kalman Filter procedure allows the aggregation of several variables with distinct frequencies for extracting the sentiment indicator and moderates the data to some extent. The analysis of micro-blog data and Twitter data may give a deeper understanding of user behavior and sentiments in the financial domain. Financial services like Thomson Reuter, Bloomberg and Eikon conduct Sentiment Analysis of Tweets, using periodicity of applied variables. Type of stock for financial analysis is common, but no one has predicted survey sentiment indices. Financial data, surveys, message boards and news outlets were some of the sources of information used to determine market sentiments and attention indicators until 2010. After 2011 The American Association of Individual Investors (AAII) and Investors Intelligence (II) were the major source to gather survey sentiments indices and attention indicators. The sentiment indicator is derived from the weekly and monthly data sets based on surveys and social media and contains interfering noise. The use of the Kalman Filter procedure allows the aggregation of several variables with distinct frequencies for extracting the sentiment indicator and moderates the data to some extent. In the present study, the blend of a daily indicator using daily data from micro-blog and diverse weekly and monthly survey sentiment indicators by applying KF procedure was used. Sentiment and attention indicators were created using Twitter which involved more than 300 million active users. Each sentiment was given a score. The negative score was given a bearish word and the positive score the bullish word. The inclusion of the number of Tweets did not improve the forecasting variables. For forecasting of trading volume, SVM method was found to be the best. The usefulness of micro-blogging data in prediction of stock market variable has been provided. It has proved an effective research in predicting impact of micro blogging

data for stock market, but it has not been able to take care of several factors such as spams, irony and sarcasm.

3.1.2 Linguistic Analysis

This study aimed to look at the regional linguistic variation across the continental United States [17]. Geotagged Tweets within the continental U.S from October 7, 2013 to October 6, 2014 were collected. Lexical alternations were then used to look at the difference in language across the U.S. A variant preference and a mean variant preference were calculated for each county and their alternation. Multiple variations of maps were created to show the use of different alternation across the continental United States. Tweets and language were analyzed at the county level. The results reflect findings of past literature research and offer new insights into linguistic variation. Unlike previous studies carried out in this area, this study takes advantage of public data. Linguistic trends were able to be found between different regions of the continent. But Twitters demographics are vastly different than the demographics of different regions. The analysis is done by selecting the Tweets which are geo-tagged. The major challenge in doing this is the lack of information, and lack of the availability of methods to analyze Tweets based on language background. Even the data which was available by research done in previous times based on language was of urban areas and not rural. More than six million Twitter users were analyzed which had a database of around eight billion words. This helped in learning about how the region is divided in the United states based on the language spoken as well as how the language change or what kind of variations language has as we move from one location to other. In the traditional method people usually had to go from place to place based on significant change language pattern, using various linguistic methods to identify the pattern and based on the data collected decided which regions have

similar patterns. Various words were compared such as mom or mother, big or large, grandfather or grandpa, you or y'all or you all or you guys, couch or sofa, etc. Users were divided based on these words; for example, if the word yall was used more in one place instead of you guys it was taken as the word of that place. To refine the results alternate words which are not used in the region are eliminated. The present method obviates the need to meet different people and visit different locations to get this type of information which was previously limited to very few people. However, it is hard to analyze and classify many messages, because they are either spam or from organizations. Abbreviations and spelling mistakes are also common on Twitter and can be difficult to analyze. Geo tagged Tweets make up only a small portion of all Tweets, while spam and non-personal Tweets are common.

The researchers used Linguistic Analysis to determine the sarcastic sentences and differentiate them with irony [18]. Tagged Tweets #irony, #sarcasm and #not were analyzed with the help of emotional and affective words. In addition, the structures of the sentences were analyzed. The #not helped in analyzing the hidden meaning of the sentence. Researches pulled these Tweets, and this helped in analyzing sentences which were sarcastic or ironic. The #not was also very useful in determining actual meaning eg. #notsarcasm is likely to depict irony. The research presents an effective way of analyzing sarcasm and irony, but since the use of language is very diverse it is hard to predict and extract sarcastic or irony filled Tweets with high accuracy.

The contents in Twitter data have a significant impact on the language of high school students. Due to the limitations of the Tweets to just 140 characters it is challenging to express a persons view and still use correct grammar. This results in the use of a lot of abbreviations and, short-forms. This can make Tweets hard to read or make sense of. There are many challenges in it and a lot of grammatical mistakes are most of the time intentionally made to express the views to the end user [19].

This is based on how peers use language on Twitter and making correct grammar a trend between the learning groups. It has been found that Twitter can be used to improve the English of the students. These conclusions were made on the basis that the simplicity of the interface is very easy to use and understand. One can relate multiple topics and give reference of different activities just by putting tags. The size of the Tweets is so small that it retains the interest of the reader. The small size also helps in the quick reply of the Tweet, as the Tweets are saved and can be accessed later and provides tractability of the progress made. These factors open doors for research in this field. To conduct the experiment a selected group of students Tweets were observed. An interactive program was prepared to analyze the grammatical change in high school students brought out by Twitter. These changes were then analyzed for the grammatical errors. These grammatical errors were provided to the moderator. With the help of moderator, a better method of implementing the same text was analyzed. This yielded positive results as the students were aware of the alternatives. It was also found that it had a huge influence on the writing skills of the peers as they tend to see how other students were using the language. The automatic text analysis has great challenges for accuracy of outcome but, the consistent efforts and improvement in the models yielded positive results. It is very effective and interesting way to improve the grammar of students.

3.1.3 Comparative Analysis

A comparison of popularity between Facebook and Twitter among college students was made using analysis including Fishers exact test, Wilcoxon matched pair sign test [20], Friedman rank-sum tests, and logistic regression [21]. The research was done on a fixed number of participants to analyze alcohol reference in their Facebook and Twitter account. It was found that college students use more alcohol references in

their Facebook than Twitter, as they feel more connected with people on Facebook. Also, it was found that the students who were putting the posts about drinking or alcohol on Twitter were people who used Twitter on regular basis.

The data from Twitter and Weibo was collected and a comparison was made based on the content, the frequency of posting and the number of times Retweeted. It was aimed at extracting information of how people of different cultural backgrounds respond to the awareness of risk and in cases of emergency [22]. Tweet Archivist was used to analyze Twitter feed and Weibo Search to analyze time of posting of the Tweet and the content in the coding itself. A sample of around 1000 posts was selected from a large pool of posts to be analyzed from each of the platform. It was found that in 799 Tweets posted 283 were Retweets whereas in the case of Weibo out of 1000 posts only 205 were Retweets. The Tweets were collected at the time of different emergency situations such as Winter Storm Nemo and, Eastern China smog. In case of Eastern China Smog, it was analyzed that most of the Tweets were at the time of the event though many continued to Tweet later. In case of analysis of text 41.0% of these Retweets were classified as relevant with the information, 34.9% were classified as a means of communicating joke, 23.0% were classified as affective, 0.7% were classified as spam, and 0.4% were categorized as insults. It helped in understanding why and how people create and consume information in crisis situations.

3.1.4 Influencer Identification

This study aimed to find common characteristics between opinion leaders on Twitter and how they may differ from the characteristics of opinion leaders in the real world [23]. A web-based survey using 648 participants at a public University lasting from October 10-25, 2014 was used to collect data. The survey used self-identification methods to assess opinion leadership. The survey asked questions about opinion lead-

ership on Twitter and offline. It was found that while gender and income had positive associations with real world opinion leaders, these characteristics had little association with opinion leaders on Twitter. Furthermore, race and household income were not associated with Twitter opinion leaders. Network size had a positive association with Twitter opinion leaders as well as offline opinion leaders. Twitter opinion leaders showed trends of civic participation online while there was no association between offline civic participation and Twitter opinion leaders. The number of Tweets and Retweets also showed a positive association with opinion leaders on Twitter. Trends between opinion leaders were clearly defined but the result is based solely on college students and self-identified persons.

3.1.5 Disaster Management

This study set out to analyze the way Twitter is used during disaster situations, specifically for Japans tsunami [24]. A narrative analysis approach was used in the study. Tweets were qualitatively categorized, and the answers obtained from questions sent out were summarized. First Tweets were gathered within a fifteen-mile radius of the two areas struck using a keyword search and then categorized qualitatively. Surverygizmo.com was used to send out an online survey to 200 people across Japan to collect more data. Most of the Tweets collected about the disaster fell into three categories: warnings, help requests, and reports about self and the environment. The major concern of Twitter users during the disaster was the reliability of the Tweets. Users could not discern true information from false information. Users also found it difficult to locate actual important messages related to the disaster. The Retweet function on Twitter caused users to see the same Tweets again and again. Highlighted users concerns about Twitter in disaster situations and showed how Twitter was used during the tsunami in Japan. The drawback of the method

is that the reliability and the number of Retweets create extra noise when searching through Tweets.

3.1.6 Cybercrime Detection

Cyberbullying has become a serious problem within the vast amounts of social networks [25]. Geo-tagged Tweets were collected within the state of California between January and February 2015. Network information, activity information, user information, and Tweet content were key features used in the machine learning algorithms to detect cyberbullying. Different combinations of the features were tested with four different classifiers to see which features and classifiers would give greater accuracy in detecting cyberbullying. Naive Bayes [26], support vector machine [27], random forest [28], and k-nearest neighbor were the machine algorithms used in this study. To determine which features gave the best results, chi squared test, information gain, and Pearson correlation selection algorithms were performed. It was found that cyberbullying victims present a higher risk of suicidal ideation. The studys goal was to be able to differentiate cyberbullying Tweets from non-cyberbullying Tweets using key features and machine learning algorithms. The machine learning algorithms could correctly label a non-cyberbullying Tweet 99.4% of the time while it could only correctly label a cyberbullying Tweet 71.4% of the time. The model in the study could be used in the feature to help with cyberbullying detection. The methods used could accurately label both non-cyberbullying Tweets and cyberbullying Tweets. The results are based on the Tweets collected within one state over a small window of time which limits the veracity of results.

3.1.7 Public Health Service

Information on the use of marijuana in different parts of America was gathered. The Tweets were filtered from the drug related Tweets using Twitters API. Keywords used for this research were: Tweets: dabs; hash oil; butane honey oil; smoke/smoking shatter; smoke/smoking budder; smoke/smoking concentrates. The eDrugTrends system provided a Twitter filtering and aggregation framework [29]. To account for different levels of Twitter activity in each state, the ratio of dabs-related Tweets to the number of general Tweets was calculated. States were classified into three categories. Status 1 indicated the state had legalized the recreational and medicinal use of cannabis. Status 2 indicated the state had legalized only the medicinal use of cannabis. Status 3 indicated that neither recreational nor medical use of cannabis was allowed in the state.

A sample of 125,255 Tweets was collected over a two-months period, out of which only 27,018 of the Tweets contained geolocation information. It was found that California, Texas, Florida, and New York had the highest raw number of dabs-related Tweets, but after adjusting for different activity levels in each state, Oregon, Colorado, and Washington had the highest proportion of dabs-related Tweets. The study found that dabs-related Tweets were more common in states where medical and recreational use of cannabis is legal. Furthermore, the Western region of the United States of America had greater dabs-related Tweeting activity. Using this method, it was able to monitor and analyze dabs-related Tweets to see dab trends. Twitter can be used in detection of drug use and its trends can be evaluated. The data collected through Twitter is quicker and more efficient than traditional drug survey methods.

This method is limited in that it can only analyze Tweets that were authored in English. The study also did not consider the demographics of the people Tweeting

the information. Researchers were not able to discern between the use of marijuana for medicinal purpose and recreational use. Also, keywords like wax were ignored in the study as they may cause inaccurate information.

3.1.8 Disease Management

This study aimed to track flu trends using Twitter and be able to predict influenza like illnesses (ILI) [30]. A Twitter Crawler was used to collect data on time intervals containing the keywords Flu, Swine Flu, and H1N1. Multiple datasets were then made from the collected Tweets. The datasets were: the original Twitter dataset, no Retweets, and no Tweets authored by the same user within 1 week, 2 weeks, and 6 weeks. An auto-regression with exogenous inputs, or ARX, was then used to create prediction models. The dataset that used no Retweets and no Tweets authored by the same user within one week provided the best results. There can be a one to two-week delay between the diagnosis of a patient and when that data becomes available on illness reports in the traditional style of doctors diagnosing and reporting the diagnosis into a system. There is a high correlation between the number of flu related Tweets and influenza like illnesses activity in the Center for Disease Control data. There was a Pearson correlation coefficient of 0.9846. A regressive model was built and tested with old CDC data. Using Twitter data improved the models accuracy in predicting ILI cases and can provide real time analysis of influenza activity. Twitter was shown to be able to effectively track influenza like illnesses and help accurately predict influenza activity.

3.1.9 Future Prediction

Twitter has come out as a wonderful source of accessing real time data generated by people. It has the advantage of following trends by looking at the previously

generated Tweets with due course of time. Therefore, it provides many opportunities in various fields for the research and is now being considered seriously [31]. Twitter's feed is analyzed for keywords based on the field of research. The data is then collected to analyze for results. Though there is no fixed way by which we can analyze the Twitter data, two approaches are discussed in this section one being quantitative and another qualitative. The way to search Twitter is to look through words or hashtag. The problem arises when we must look for general things. This means it is hard to figure out any general trend or topics as there are no specific keyword to determine all the related Tweets; whereas it is easy to find out any specific information by putting in keywords. To analyze these Tweets various graphs are drawn and analyzed in the background. Multiple logical connections are made on several factors based on the kind of analysis to form relationship between forms of data and the useful results obtained from them. It is an effective way of searching with the help of keywords. Dependence on keywords, and lack of understanding language features such as sarcasm are the major limitations.

3.1.10 Medical Complaints

It is well known that medical errors lead to many serious health issues. To understand these errors and their impact in society we need a feedback system. The conventional methods have not proven very effective as many people do not provide feedback or are aware of what they are going through because of the side effects of certain medicine. With the rapid increase in use of the internet and awareness between people analysts have started using social media to take feedback [32]. Patient's safety related Tweets were identified and classified based on self-reports or those reported by others. A pattern of key words was formed for example keywords such as, sue hospital, nurse, or doctor paired with terms such as screwed up, fucked up, mistake, was wrong, etc.

Here we are emphasizing how the use of Twitter is being made to get instantaneous and valuable feedback. The criteria of this research is to identify the patients, who are providing the feedback and what kind of errors are being reported. This in due course of time will be helpful in tackling as well as minimizing these issues. To study this problem, a Tweet set of one year was collected with the help of experts and analyzed for keywords which included medical related terms, common side effects, use of certain terms which connotes anger or frustration, and the use of terms referring to the hospital or the staff members, such as doctor, nurse, surgeon, etc. They were then classified based on the seriousness of the issues and side effects. It was found that most of these Tweets were made by the patients themselves and others were made either by their relatives or from unknown sources.

3.1.11 Computer Communications

The Soft Frequent Pattern Mining (SFPM) based approach was used in this study [33] to detect relevant topics on Twitter. The SFPM was adjusted to meet the constraints of a dynamic detection scenario. Using the Term Frequency Inverse Document Frequency allows for filtering of common terms and for selecting relevant terms.

In the study, researchers used the SFPM based approach to detect current, relevant topics being discussed on Twitter. Data was collected, and tests were run during the 2014 FIFA World Cup using the modified SFPM approach and a basic SFPM approach. The experiments used five matrices, which were topic recall, keyword precision, keyword recall, precision and redundancy. The modified SFMP in the study outperformed the basic SFMP approach in detecting relevant topics. The system used in this study was designed to adapt its behavior based on the incoming Tweets. The SFPM algorithm has superior performance in detecting relevant topics when compared to other techniques. The SFPM allows users to easily change and

update the way in which current terms are picked up while still being efficient. More work needs to be done in summarization techniques of texts to speed up the evaluation process.

3.2 Review of the Results

Figure 3.1 shows the screening of the research papers.

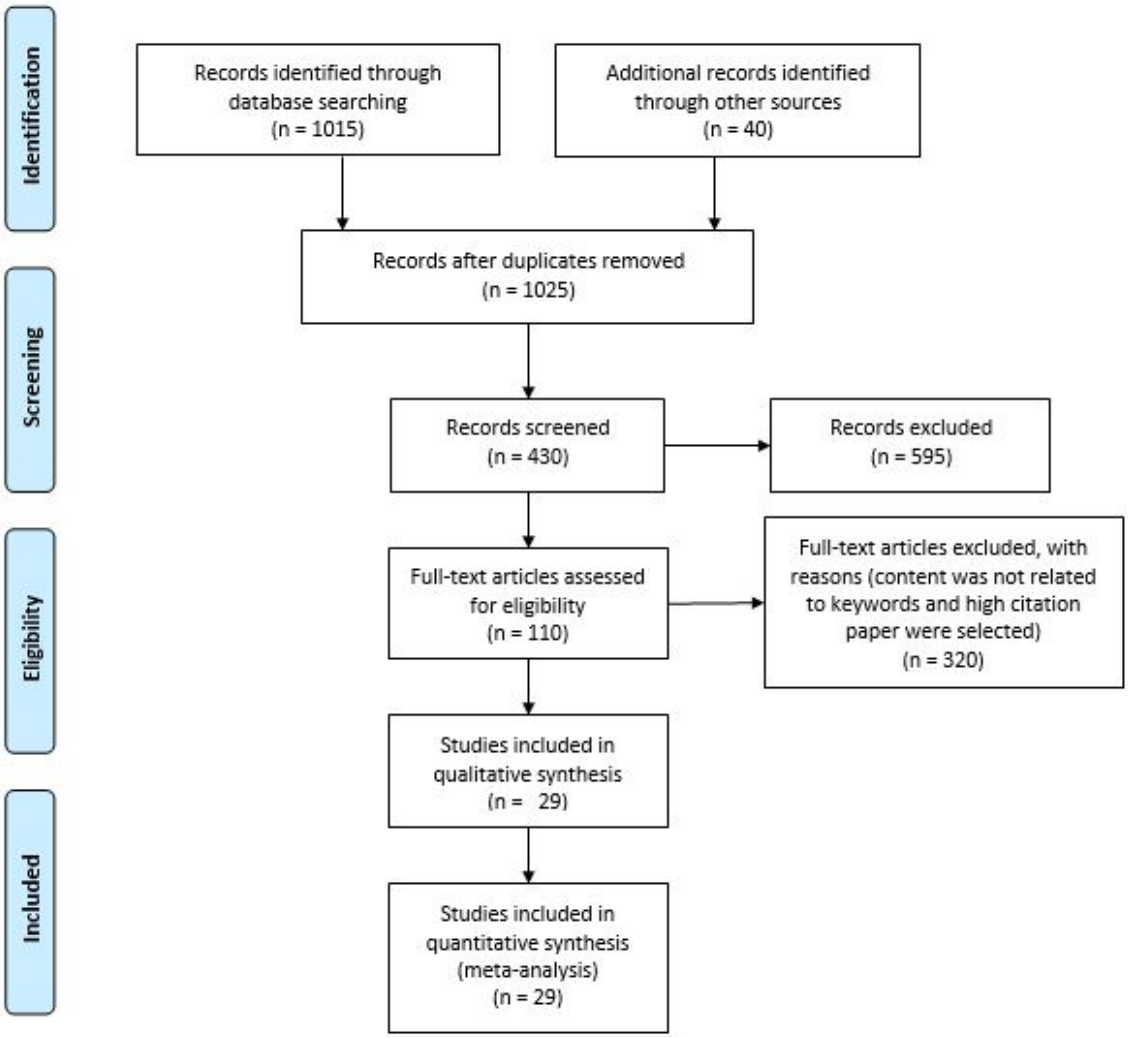

Figure 3.1: Flow diagram showing screening of the research papers

Papers are selected based on field of research, year of publication and journals they are published in. They are represented in Tables 3.1, 3.2, and 3.3 respectively.

Field of research	Available papers with keywords	Selected based on relevance
Sentiment Analysis	261	10
Linguistic Analysis	150	3
Comparison of Data	75	2
Influencer Identification	54	2
Disaster Management Processing	63	2
Information System	34	2
Impact on People	21	1
Detecting Relevant Topic	4	1
Computer Communication	5	1
Cybercrime Detection	68	1
Public Health Service	120	1
Disease Management	78	1
Future Event Prediction	49	1
Medical Complaints	43	1
Total	1025	29

Table 3.1: Selection of papers based on field of research

Year	Number of Research Paper
2011	1
2012	1
2013	0
2014	0
2015	6
2016	16
2017	5

Table 3.2: Selection of papers based on year of publication

Name of Journal	Amount	Percentage
Computers in Human Behavior	6	20.69
Information Processing and Management	2	6.89
Computers, Environment and Urban Systems	1	3.45
Knowledge-Based Systems	3	10.34
Expert Systems with Applications	2	6.89
Transportation Research Part C: Emerging Technologies	1	3.45
Technological Forecasting and Social Change	1	3.45
Internet Interventions	1	3.45
Journal of Adolescent Health	1	3.45
Drug and Alcohol Dependence	1	3.45
Computer Communications	1	3.45
Journal of Surgical Research	1	3.45
Online Social Networks and Media	1	3.45
Big Data Research	1	3.45
Safety Science	1	3.45
Computer Communications Workshops	1	3.45
Decision Support Systems	1	3.45
International Journal of Web Based Communities	1	3.45
Futures	1	3.45
Engineering Applications of Artificial Intelligence	1	3.45
Engineering Applications of Artificial Intelligence	1	3.45
Total	29	100

Table 3.3: Selection of papers based on journal

Chapter 4: Development of Sentiment Analysis Framework and Case Study

4.1 Twitter Sentiment Analysis Framework

Various tools and techniques have been used in the working of the framework and verification of it. Some of the technologies are readily available to be utilized for different purpose and proper working of the application. The tools used in this application are explained in the subsequent sections.

4.1.1 Description of Tools used in the Framework

Django

Django [34] is a Python based, high level web framework. It is designed by highly skilled and experienced developers. Django helps in rapid development which is sensible and realistic rather than theoretical consideration. Therefore, developer are able to focus more on application rather than making the web design work. Another advantage is that it is free and open source. Key features of Django that makes it ideal for this application include: it allows rapid development, as it was designed with the motive of taking application from concept to execution very quickly; it is very secure; it is highly scalable, hence many busy websites such as bitbucket use it for quick and flexible usage.

Amazon Web Services (AWS)

Amazon Web Services (AWS) [35] is a subsidiary of Amazon.com. It provides hardware services such as CPU(s) and GPU(s) for processing, hard-disk/SSD storage, local/RAM memory, it also provides a choice of operating systems, networking, and pre-loaded application software such as web servers, databases, CRM, etc. It provides services to more than 1 million customers such as Spotify, Airbnb, Shazam, Comcast and Johnson and Johnson.

Our application requires hosting, to make the application usable from any location, and to help access the data remotely we figured that AWS was the best solution. In our application we have used a free and limited version of Amazon Web Services which provide free access to 20 GB HDD ,1 GB RAM.

Tweepy

Tweepy [36] is a Python library for accessing Tweets recommended by Twitter. Twitter recommends few library if you are using Python. Twitter requires all requests to use OAuth for authentication. Application Programming Interface (API) provides RESTful API methods. When an API method is invoked a tweepy model class instance is returned, which we use for our application. Tweepy also helps in making OAuth authentication easier. When a new application is created a customer token and a secret is generated. Then an Auth handler instance is created by passing this customer token and the secret.

TextBlob

TexBlob [37] is a Python library for processing data. In our application we have use TextBlob for Sentiment Analysis of the Tweets which is well compatible with

Python and one of the finest free tool available on the web. Python have various library for Sentiment Analysis, TextBlob has been reviewed most.TextBlob combines the PatternAnalyzer (based on the pattern library) and NaiveBayesAnalyzer (an NLTK classifier trained on a movie reviews corpus), which makes it one of the best Sentiment Analysis tool. For analyzing words and sentences properties TextBlob uses Tokenizer, which breaks the sentence in diffrent parts as shown in Figure 4.1. It uses textblob.tokenizers.WordTokenizer and textblob.tokenizers.SentenceTokenizer classes,to break words and sentence respectively.

Based on these analysis the sentence is categorized into positive, negative or neutral on its basis and the polarity of the sentence is given.

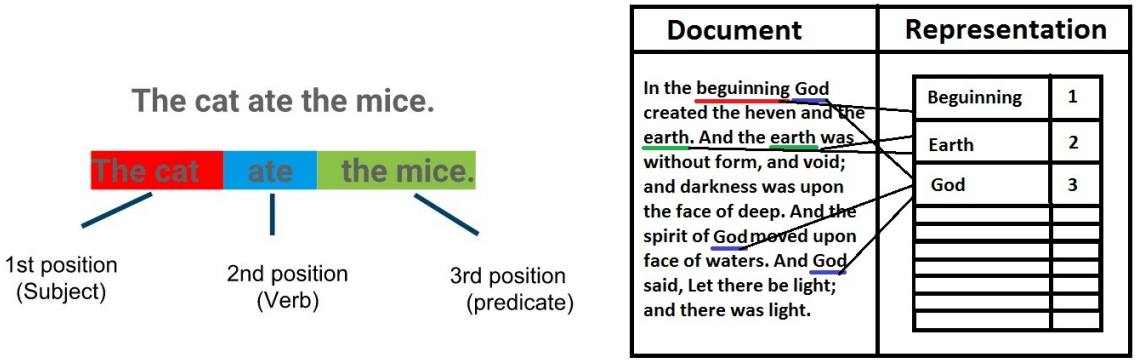

Figure 4.1: Tokenization of words in sentences

4.1.2 Verification of the Framework

Precision and recall [38] is a method of measuring the success of a prediction. Precision defines how relevant a result is, while a recall is a measurement of how many truly relevant Tweets are retrieved. The perfect precision score is 1.0 in an information retrieval this means that every Tweet retrieved by the system is relevant. On the other hand if recall has the score of 1.0, it means all the relevant Tweets have been retrieved, but it does not describe how many irrelevant Tweets were retrieved. Precision (P)

is defined as the number of True Positives (Tp) over the number of True Positives plus the number of False Positives (Fp). Recall (R) is defined as the number of True Positives (Tp) over the number of True Positives plus the number of False Negatives (Fn). In this research we have used this method to manually analyze +ve and -ve sentiment Tweets to check the authenticity of the Sentiment Analysis tool on which the reliability of the application is dependent.

4.1.3 Framework

The Twitter Sentiment Analysis Framework as shown in Figure 4.2 uses Tweepy, which is a Python library that collects Tweets from Twitter using customer token and the secret. The collected Tweets are analyzed for sentiments with help of TextBlob which feeds it to Django view for the representation of data and Django model for storing data into SQLite for later use.

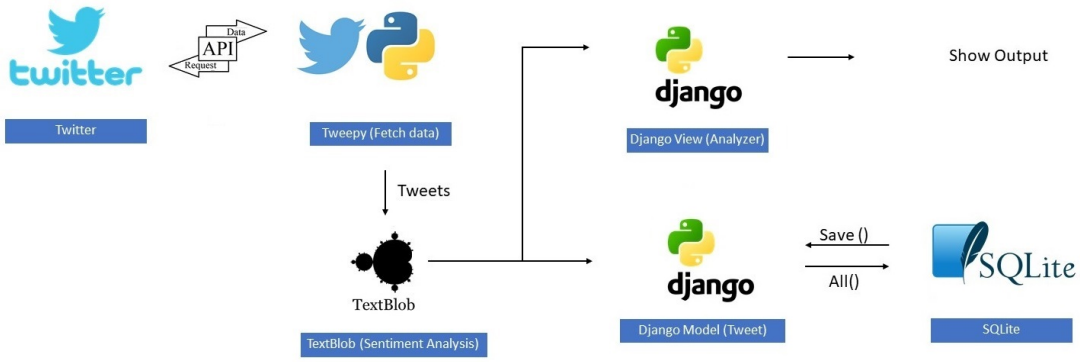

Figure 4.2: Framework of the application

Key features of our framework are:

Our Framework give graphical representation in donut graph and represents the polarity of Tweets. It is user friendly and can be reused for other parameters. Most

of the other tools are designed to tackle a specific problem. Many frameworks require knowledge of Python or other programming languages to use. The framework also stores data in the application itself and displays it there and not in a separate database. It stores the data in three categories, positive, negative and neutral.

4.1.4 Data Extraction

Twitter provides limited access of the data through the Twitter API. The application uses a Python-based library Tweepy to access Tweets through the Twitter. To access Tweets the user should have a Twitter account. Twitter requires all requests to use OAuth for authentication. Application Programming Interface (API) provides RESTful API methods. When an API method is invoked a Tweepy model class instance is returned, which we use for our application. Tweepy also helps in making OAuth authentication easier. When a new application is created a customer token and a secret is generated. Then an auth handler instance is created by passing this customer token and the secret.

4.1.5 Sentiment Analysis

In our application we have used TextBlob because it is compatible with Python. It combines two approaches for working: the bag of words approaches and the lexicon-based approach. In the bag of words approach, it divides the sentence into various parts and collects the word counts. The second is the lexicon-based approach, in which the words are compared with pre-stored words based on sentiments. The words in the sentence are compared and the sentence is categorized into positive, negative or neutral on this basis. The TextBlob collects the Tweets fetched from the Tweepy library and performs the Sentiment Analysis. It then sends the data to Django view and Django model so that it can be represented in a user friendly way and can be

stored in database for later use.

4.1.6 Data Visualization

The framework in Figure 4.3 displays the keywords searched, the number of Tweets fetched and the polarity of these Tweets. The Sentiment Analysis of Tweets is presented in the doughnut chart. It also displays text of the Tweets in three categories positive, negative and neutral.

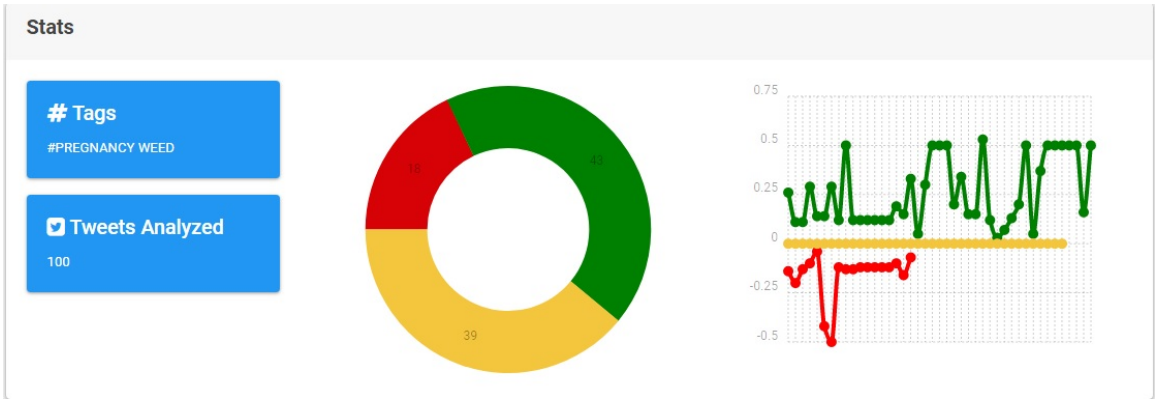

Figure 4.3: Visualization of data

4.2 Case Study

Keywords for the research were selected based on their rapid mentions in Tweets and on the internet. We collected 31 keywords for pregnant women and 30 for marijuana. We tried different combinations of these keywords to get the maximum relevant output and maximum number of Tweets. The best combination of keywords of pregnant and marijuana were selected to be #Pregnant and #Marijuana, #Pregnancy and #Marijuana, #Pregnant and #Weed, #Pregnancy and #Weed, #Pregnant and #Cannabis, #Pregnancy and #Cannabis. This combination is also a good way of showing how the sentiments of people changes when they talk about pregnant women

and marijuana and when they talk about marijuana and pregnancy. The other keyword combinations fetched very few or no Tweets for example

#Pregnancylife and #pot gave 1 result #BabyOnTheWay and #420 gave 0 result #PREGNANTLIFE and #GANJA gave 0 results.

So we selected the most common words used to represent pregnancy and marijuana to increase the number of results and related Tweets. The selected keywords gave 70 to 100 results per search.By using these keywords I was able to identify how the sentiments of people change when talking about e.g. pregnant and pregnancy.

We collected 19116 Tweets based on these keywords. These Tweets were analyzed by our framework and were divided into three categories positive, negative and neutral. The readings were taken 3 times a day from the 16th of March 2018 to 8th of April 2018. The averages of these sentiments were computed based on the selected keywords and put in the graph as shown in Figure 4.4.

As our tool can store positive, negative and neutral Tweets separately, We took a sample of 300 Tweets for manual analysis. From these Tweets 150 Tweets were positive sentiments and 150 Tweets were negative sentiments. Neutral sentiments Tweets were not considered in our analysis. The positive and negative sentiment Tweets were read critically and based on the sentiments expressed in them they were judged to have been classified or misclassified by the tool as, True Positive (Tp), False Positive (Fn), True Negative (Tn) and False Negative(Fn). A False Positive label indicates that it was determined that the tool incorrectly classified a negative Tweet as positive, and a False Negative indicates that the tool incorrectly classified a positive Tweet as negative. For further research we also categorized these Tweets based on gender, i.e. if user is likely to be a male (Likely Male), if user is likely to be a female (Likely Female) and or if user is unknown (Unknown). Further we also tried to categorize, if the Tweet is an Opinion, Information or Marketing Tweet.

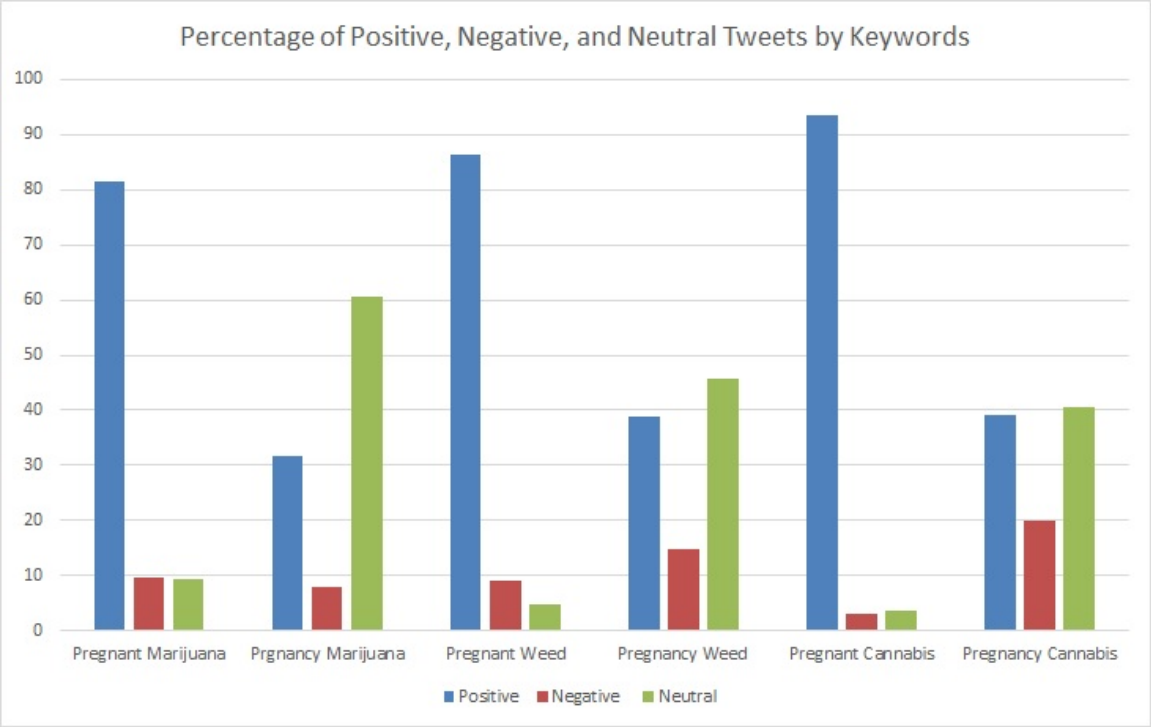

Figure 4.4: Sentiments based on keywords

True Positive Tweets and False Positive Tweets were counted and came to 106 and 44 respectively. Similarly, True Negative Tweets and False Negative Tweets were counted and came to 106 and 44 respectively. This data is given in Table 4.1 and is represented in Figures 4.5 and 4.6.

Sentiments	Number of Tweets	True	False
Positive	150	106	44
Negative	150	112	38

Table 4.1: Manually analyzed tweets

The False Positive Tweets were in fact negative Tweets and the False Negative Tweets were in fact positive Tweets. The data was then subjected to information

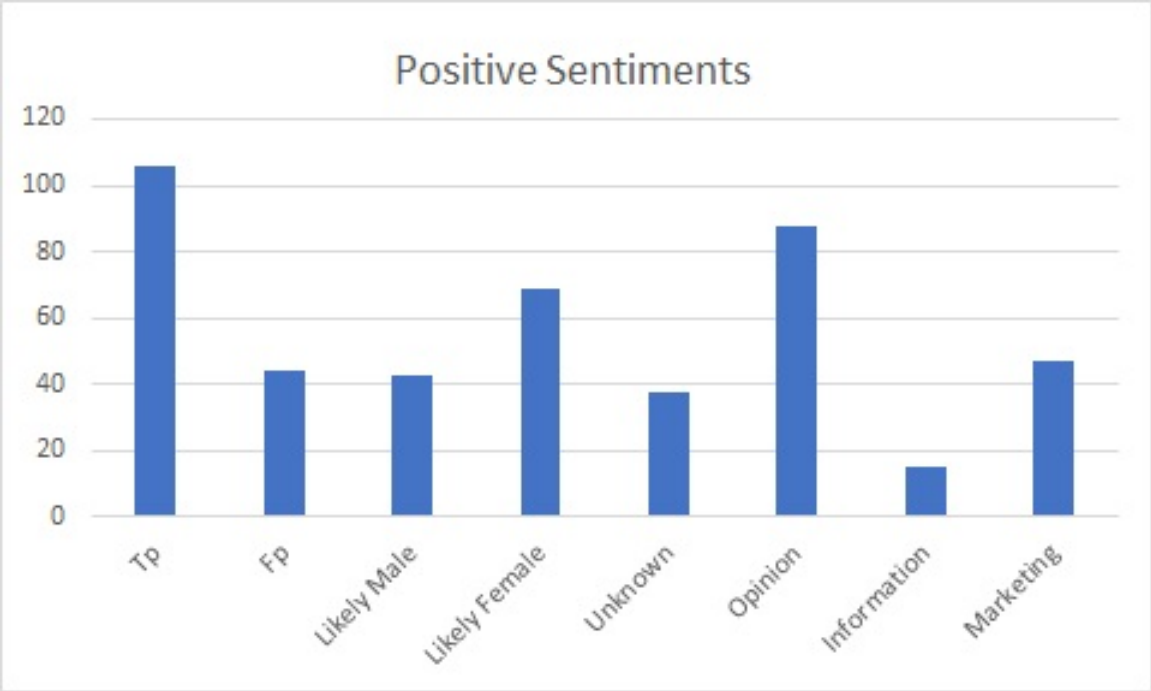

Figure 4.5: Positive sentiments

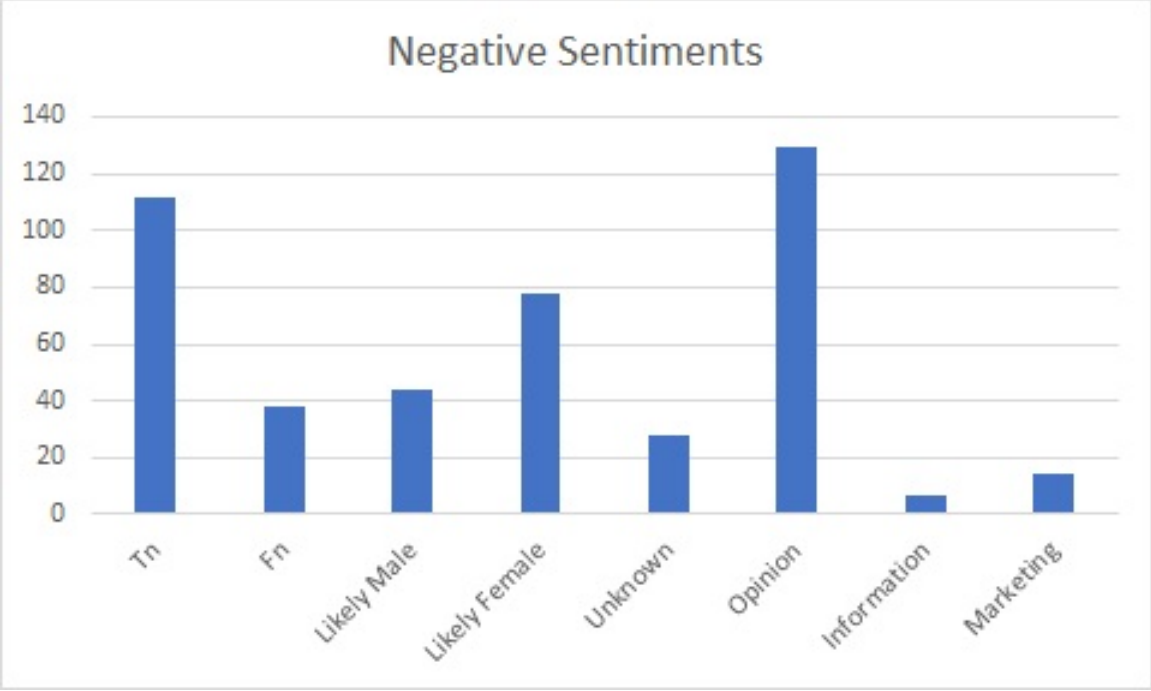

Figure 4.6: Negative sentiments

retrieval, by Precision and Recall the validity of the framework was established. Precision is a fraction of relevant information among the retrieved information. Precision for the positive Tweets was determined by using the formula, shown in Equation 4.1.

$$Precision = Tp/(Tp + Fp) \tag{4.1}$$

Where Tp = True Positive and Fp = False Positive

Precision = 106/ (106+44) = 0.707. This is also called Positive Predictive Value (PPV)

Likewise the Negative Predictive Value (NPV) or precision for negative Tweets is calculated by the formula shown in Equation 4.2.

$$NegativePredictiveValue = Tn/(Tn + Fn) \tag{4.2}$$

Whereas Tn = True Negative and Fn = False Negative

Negative Predictive Value = 112/ (112+38)= 0.747

Recall or the sensitivity is a fraction of information that have been retrieved out of total amount of relevant information available.

Recall or sensitivity of the framework was determined using the formula, shown in Equation 4.3.

$$Sensitivity = Tp/(Tp + Fn) \tag{4.3}$$

Recall/ Sensitivity = 106/ (106+38) = 0.736

Though there is inverse relation between precision and recall, we obtained 0.707 and 0.747 values for precision for positive and negative retrieved information respectively; whereas for recall the value is 0.736 which is fairly high. True Negative Rate

(TNR) which is also called Specificity is determined by Equation 4.4.

$$Specificity = Tn/(Tn + Fp) \tag{4.4}$$

Specificity or TNR = 112/(112+44)=0.718

Further the accuracy of the framework is calculated by the equation Equation 4.5.

$$Accuracy = (Tp + Tn)/(Tp + Tn + Fp + Fn) \tag{4.5}$$

Accuracy = (106+112)/ (106+112+44+38) = 0.727

The Predictive Positive Condition Rate (PPCR), shown in Equation 4.6 is the percentage of positive expected returns out of total population.

$$\{(Tp + Fp)/(Tp + Fp + Tn + Fn)\}100 \tag{4.6}$$

Substituting the values of Tp, Fp, Tn and Fn one can achieve

{(106+44)/(106+44+112+38)}100 = 50%

The combination of Precision and Recall is the harmonic mean of precision and recall and is denoted by F-measure, shown in Equation 4.7.

$$F = 2(PPV * TPR)/(PPV + TPR) \tag{4.7}$$

Substituting values of PPV,TRP one can achieve

F = 2(0.707 x 0.736)/ (0.707+0.736) = 0.721

The measure is approximately the average of the two when they are close and is called the harmonic mean. In our case the average of Precision and Sensitivity is

(0.707+0.736)/2 = 0.722 which is very close to the harmonic mean 0.721.

The gender of the Tweeters were identified based on their name, photo and the

language of the text. In a few cases it was not possible to determine gender, which were placed in the category of unknown. The male/female ratio of positive sentiment Twitters was 1:1.605 and negative sentiment Twitter accounts was 1:1.773 and in 66 cases it was not possible to find out their gender. The Tweets were collected at 11 am, 6 pm and 11 pm with the understanding that some people would like to Tweet in the morning, others in evening or at the end of day before retiring on bed, which might affect the type of Tweets but we did not find any significant difference in the number or type of Tweets based on the time of collection.

The personal opinion or the Tweets in which person express his or her own opinion on positive sentiment was expressed by 88 Twitters and negative sentiments by 129 Twitters. The positive sentiments on marketing which are the Tweets by company or organizations were given by 47 Twitters and negative sentiments by 14 Twitters.

Chapter 5: Conclusion and Future Work

The aim of this systematic review was to study the various developments in the field of Twitter data analytics and understand their applications and methods. For better understanding, we describe the idea of the review methods and the various steps involved. Based on these principles, various research papers were analyzed and classified into four categories based on selected parameters. The classification is helpful for any person who wants to understand the basics of Big Data methods of collecting, processing and finding insights out of the data collected. We also described various ongoing research in the field of Twitter data extraction which utilizes the concept of the data collection by means of Big Data. This helps in understanding the ongoing research in field of Twitter data analytics. This thesis does not claim to be comprehensive, but we have tried to put some ongoing research in the field of the Twitter data analysis. With rapid advancements in the field of Twitter data analytics, we recommend re-visiting these methods and periodically revising the review to include new developments.

The analysis shows that most of the research papers were based on the Sentiment Analysis. Twitter data has proved itself as a good way of analyzing the sentiments of large group of people. The case study conducted in this research also helps in better understanding of the usage of Twitter data and serves as an example to give a good idea of the kind of analysis.

The framework introduced for sentiment analytics of pregnant women on mari-

juana using Twitter data was analyzed on 19116 Tweets. Out of all the keywords used, the best results were achieved by the following combination of keywords: pregnant and marijuana, pregnant and weed, pregnant and cannabis, pregnancy and marijuana, pregnancy and weed, pregnancy and cannabis. The framework returned 60-100 Tweets which amounted to 360-1200 Tweets per search per day.

The sentiments vary tremendously when different terms were used, e.g. pregnant and cannabis gave result of an average ranging from 89 to 100% positive, 0 to 5% negative and 0 to 6% neutral. When we changed the terms to pregnancy and cannabis the results changed drastically; we see only 34 to 49% positive sentiments, 0 to less than 2% negative sentiments and a large number of neutral sentiments. On analyzing the Tweets, we discovered the difference that when individuals sentiments are associated with the word pregnant and cannabis, they tend to be very positive about it but when they talk about pregnancy they tend to be conscious and have a lot of questions. This was also confirmed when we compare the results of searches with the keywords pregnant and marijuana, pregnancy and marijuana as well as pregnant and weed, pregnancy and weed. Most of the Tweets were mainly about their opinion on this topic, or on the information they wish to share. Twitter accounts showed most of the positive Tweets with the term pregnant and displayed more concern about words used with pregnancy, and were seeking for information on it. Therefore, most of the neutral Tweets were posted both by males and females in reference to pregnancy. The personal opinion was expressed by 88 positive sentiment Twitters and 129 negative Twitters while marketing sentiments were given by 47 and 14 positive and negative sentiment Twitters respectively.

Looking at the number of Tweets we determined there was no significant difference in the number of Tweets posted at different time of the day. On analysis of 300 Tweets, we calculated Precision of positive and negative sentiments. The Predictive Positive

Value was 0.707 and the Negative Predictive Value was 0.747 which is considered quite good. Recall/Sensitivity was 0.736. The True Negative Rate (TNR) or Specificity was 0.718 and the accuracy, 0.59. The F-measure (Harmonic mean) was calculated to be 0.721 which was close to 0.7215, the average of Precision and Recall. The male/female ratio of positive sentiment Twitters was 1:1.605 and negative sentiment Twitters was 1:1.773 suggesting that majority of the Tweets were made by women. From the results of this study we determined that the analytics of Tweets on important issues can be done with fair accuracy and meaningful information about the population can be obtained. A majority of people now get news via social media and more than half of the public in metropolitan areas have turned to social media sites [39]. A survey of 1522 adults in March-April 2016 finds that around 24 percent youth uses Twitter.

In future, the application can be used to collect sentiments of people about any issue being discussed on Twitter. Some Tweets also provide geo-code, through which the location of people Tweeting can be traced. This will help in analyzing the sentiments of people from various locations. However, when the location of the Tweet was on by default less than one percent of the Tweets were geo-tagged. The Twitter provide only a small fraction of all the Tweets available for extraction of data, which by itself reduces the chances of people sharing location. Since now the location is by default closed and the topic of research is very sensitive, it is highly unlikely to fetch geo-tagged Tweets. We tried to collect Tweets by fixing the coordinates and looking for Tweet around the area of 1500 kilometers and was not able to retrieve single Tweet on this topic. Recently, the privacy policy of users have been updated and users have to manually turn on their location now, which is done by a very small fraction of society. Especially as very few people want to share their location

BIBLIOGRAPHY

[1] https://www.sas.com/en_us/insights/analytics/big-data-analytics.html.

[2] https://www.slideshare.net/BernardMarr/140228-big-data-slide-share/10-With_the_datafication_comesbig_data.

[3] https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care.

[4] https://www.slideshare.net/BernardMarr/big-data-25-facts.

[5] https://www.lexalytics.com/technology/sentiment.

[6] M. Birjali, A. Beni-Hssane, and M. Erritali, "Analyzing social media through big data using infosphere biginsights and apache flume," *Procedia Computer Science*, vol. 113, pp. 280–285, 2017.

[7] J. P. Carvalho, H. Rosa, G. Brogueira, and F. Batista, "Misnis: An intelligent platform for twitter topic mining," *Expert Systems with Applications*, vol. 89, pp. 374–388, 2017.

[8] X. Xiao, A. Attanasio, S. Chiusano, and T. Cerquitelli, "Twitter data laid almost bare: An insightful exploratory analyser," *Expert Systems with Applications*, vol. 90, pp. 501–517, 2017.

[9] A. O. Durahim and M. Coşkun, "# iamhappybecause: Gross national happiness through twitter analysis and big data," *Technological Forecasting and Social Change*, vol. 99, pp. 92–105, 2015.

[10] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015.

[11] Y. Yu and X. Wang, "World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans tweets," *Computers in Human Behavior*, vol. 48, pp. 392–400, 2015.

[12] F. Corea, "Can twitter proxy the investors' sentiment? the case for the technology sector," *Big Data Research*, vol. 4, pp. 70–74, 2016.

[13] https://klout.com/corp/score.

[14] M. Daniel, R. F. Neves, and N. Horta, "Company event popularity for financial markets using twitter and sentiment analysis," *Expert Systems with Applications*, vol. 71, pp. 111–124, 2017.

[15] A. C. Pandey, D. S. Rajpoot, and M. Saraswat, "Twitter sentiment analysis using hybrid cuckoo search method," *Information Processing & Management*, vol. 53, no. 4, pp. 764–779, 2017.

[16] N. Oliveira, P. Cortez, and N. Areal, "The impact of microblogging data for stock market prediction: using twitter to predict returns, volatility, trading volume and survey sentiment indices," *Expert Systems with Applications*, vol. 73, pp. 125–144, 2017.

[17] Y. Huang, D. Guo, A. Kasakoff, and J. Grieve, "Understanding us regional linguistic variation with twitter data analysis," *Computers, Environment and Urban Systems*, vol. 59, pp. 244–255, 2016.

[18] E. Sulis, D. I. H. Farías, P. Rosso, V. Patti, and G. Ruffo, "Figurative messages and affect in twitter: Differences between# irony,# sarcasm and# not," *Knowledge-Based Systems*, vol. 108, pp. 132–143, 2016.

[19] M. Oussalah, B. Escallier, and D. Daher, "An automated system for grammatical analysis of twitter messages. a learning task application," *Knowledge-Based Systems*, vol. 101, pp. 31–47, 2016.

[20] https://statistics.laerd.com/spss-tutorials/wilcoxon-signed-rank-test-using-spss-statistics.php.

[21] M. A. Moreno, A. Arseniev-Koehler, D. Litt, and D. Christakis, "Evaluating college students' displayed alcohol references on facebook and twitter," *Journal of Adolescent Health*, vol. 58, no. 5, pp. 527–532, 2016.

[22] X. Lin, K. A. Lachlan, and P. R. Spence, "Exploring extreme events on social media: A comparison of user reposting/retweeting behaviors on twitter and weibo," *Computers in Human Behavior*, vol. 65, pp. 576–581, 2016.

[23] C. S. Park and B. K. Kaye, "The tweet goes on: Interconnection of twitter opinion leadership, network size, and civic engagement," *Computers in Human Behavior*, vol. 69, pp. 174–180, 2017.

[24] A. Acar and Y. Muraki, "Twitter for crisis communication: lessons learned from japan's tsunami disaster," *International Journal of Web Based Communities*, vol. 7, no. 3, pp. 392–402, 2011.

[25] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.

[26] NaiveBayesian.http://www.statsoft.com/Textbook/Naive-Bayes-Classifier.

[27] http://scikit-learn.org/stable/modules/svm.html.

[28] RandomForest.http://www.stat.berkeley.edu/~breiman/RandomForest/cc_home.htm.

[29] R. Daniulaityte, R. W. Nahhas, S. Wijeratne, R. G. Carlson, F. R. Lamy, S. S. Martins, E. W. Boyer, G. A. Smith, and A. Sheth, "time for dabs: Analyzing twitter data on marijuana concentrates across the us," *Drug & Alcohol Dependence*, vol. 155, pp. 307–311, 2015.

[30] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," in *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*. IEEE, 2011, pp. 702–707.

[31] V. Kayser and A. Bierwisch, "Using twitter for foresight: An opportunity?" *Futures*, vol. 84, pp. 50–63, 2016.

[32] A. Nakhasi, R. Passarella, S. G. Bell, M. J. Paul, M. Dredze, and P. Pronovost, "Malpractice and malcontent: Analyzing medical complaints in twitter," in *2012 AAAI Fall Symposium Series*, 2012.

[33] S. Gaglio, G. L. Re, and M. Morana, "A framework for real-time twitter data analysis," *Computer Communications*, vol. 73, pp. 236–242, 2016.

[34] https://www.djangoproject.com/.

[35] https://aws.amazon.com/.

[36] http://www.tweepy.org/.

[37] http://textblob.readthedocs.io/en/dev/.

[38] http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html.

[39] http://www.pewinternet.org/2016/11/11/social-media-update-2016/.