TWITTER SENTIMENT ANALYSIS TO STUDY ASSOCIATION BETWEEN FOOD HABIT AND DIABETES

A Thesis by

Nazila Massoudian

Bachelor of Science, Tehran Azad University of Engineering and Technology, 2006

Submitted to the Department of Electrical Engineering and Computer Science

and the faculty of the Graduate School of

Wichita State University

in partial fulfillment of

the requirements for the degree of

Master of Science

May 2016

TWITTER SENTIMENT ANALYSIS TO STUDY ASSOCIATION BETWEEN FOOD HABIT AND DIABETES

The following faculty members have examined the final copy of this thesis for form and content, and recommend that it be accepted in partial fulfillment of the requirement for the degree of Master of Science with a major in Computer Science.

_____

Kaushik Sinha, Committee Chair

_____

Vinod Nambodiri, Committee Member

_____

Hamid M. Lankarani, Committee Member

DEDICATION

To my husband, my parents and my family

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Kaushik Sinha for all his guidance and support. I would also like to thank Kian for their thought provoking inputs. I would like to thank Ahmad, Mehrdad and Sara for all their helps and supports. Lastly, I would like to thank all who have ever given even a single input that in anyway helped in completing this work.

ABSTRACT

Social media platforms such as Twitter, Facebook, Instagram and etc. are rapidly becoming key resources for many researches. Among all micro blogging services, Twitter is one of the most important ones. Vast amounts of freely available, user-generated online content, in addition to allowing for efficient and potentially automated, real-time monitoring of public sentiment, allow for bottom-up discovery of emergent patterns that may not be readily detectable using traditional surveillance methodologies such as pre-formulated surveys.

One of the most significant health issues in the world and particularly in the US is the high rate of diabetes, which causes early death, cardiovascular disease and many other health problems. In this work, a framework is developed to study the association between social media attitude towards “Fast food” and reported diabetes rate available from Government websites. In this work, two classification methods are used for predicting the sentiments in tweets containing the word “Fast Food”. First method is a generic classifier that uses a predefined dictionary to compute the polarity of a given tweet due to get its sentiment; the polarity score is a float within the range [-1.0, 1.0]. Therefore, if the polarity is less than 0 the result of sentiment will be negative; if it is equal to 0, the result of sentiment will be neutral; otherwise it will be positive [1]. The second one is manually labeled classifier that uses a manually labeled training set specifically suited for the purpose of predicting tweet sentiments containing the word “Fast Food”. For both the classifiers, correlation coefficients between predicted negative tweets percentage and reported diabetes rates were computed. It was observed that negative sentiments predicted by the manually labeled classifier showed stronger correlation to the reported diabetes rate for 14 states of the U.S. as compared to that of the generic classifier.

TABLE OF CONTENT

CHAPTER 1

INTRODUCTION

Social media is part of the web in which billions of people are active and interact on a regular basis. Such user activities and public interactions posted on various social media websites, in the form of Facebook posts or tweets, when analyzed properly, provide a massive source of useful information. Therefore, the data provided by social media such as Twitter, Facebook, Instagram, etc. is highly valuable for researchers in understanding individual and social behaviors. For example social media has been used to predict movies box-office revenue [2], predict flu trends [3] and predict age of smoking [4].

Moreover the interpretation of these data can be applied to the field of Medicine and Public Health as well where it can help in deciphering the social and behavioral aspect of various diseases as well as in formulating varied intervention methods in attenuating the disease. For example, De La et al. [5] conducted a study to identify the usefulness of Facebook and Twitter in relation to groups with Colorectal Cancer (CRC), Breast Cancer and Diabetes Mellitus. They conducted a search in Facebook and Twitter using the term "colorectal cancer", "breast cancer" and "diabetes" and identified 171 colorectal cancer group, 216 breast cancer group and 527 diabetes group. Moreover they also observed that Facebook has more social groups for diabetes and breast cancer (82%) and colorectal cancer (62.23%) compared to twitter (18% for diabetes and breast cancer and 31.76% for colorectal cancer). Thus the study shows that Facebook has higher usage for disease support purposes. In another recent study conducted by Eichstaedt et al. (2015) [6] the authors used language expressed in twitter to describe the “community-level psychological correlates of age-adjusted mortality from atherosclerotic heart disease (AHD) ".

They identified that the language pattern showing negative relationships; anger; sadness are risk factors for AHD whereas as positive emotions such a happiness etc. acts a protective factor against AHD. They also conducted a "cross-sectional regression model based analysis" on twitter language which reflected AHD mortality better than a model combining the common demographic, socioeconomic and health factors such as smoking, obesity and diabetes.

1.1 Research Necessity and Objective

This thesis analyzes vast amount of user activity data available from twitter in the form of tweets and aims to discover how bad food habit pattern of people are associated with reported diabetes rates in different states of the United States. Diabetes is one of the most notorious diseases that affect millions of people from across the globe as well as from the U.S. Diabetes is a disease that causes high blood sugar. It has been categorized in two types. If a person's body could not produce enough insulin it is called diabetes type one and if a person's body could produce enough insulin but it could not use its own insulin effectively it is called type two. If someone affected with diabetes could not control this situation by injecting insulin or taking pills, diabetes could lead to serious health complications and most likely death. A risk of death for a person who has diabetes is twice the risk of a person of similar age who does not have diabetes. Diabetes is a major cause of heart disease and stroke. Moreover, 67% of U.S. adults who report having diabetes also report having high blood pressure, and as Fig 1.1 shows the number of people who have been diagnosed with diabetes is increasing every single year [7]. Therefore, to discover and study various factors that cause and/or are associated with diabetes are extremely important as they might help to pre-diagnose and prevent diabetes. In this study the percentage of US. Adults with diagnosed diabetes from14 randomly chosen states from the U.S. as shown in Table 1 [7] have been used.

Table 1-1

Percentage of adults diagnosed with Diabetes per state in 2012

State	AL	TN	GA	OH	IN	AR	PA	FL	NY	NJ	IL	AZ	WA	OR
Percentage	11.1	10.2	9.8	9.4	9.1	9.2	8.7	8.7	8.4	8.3	8.2	8.1	7.4	7.2

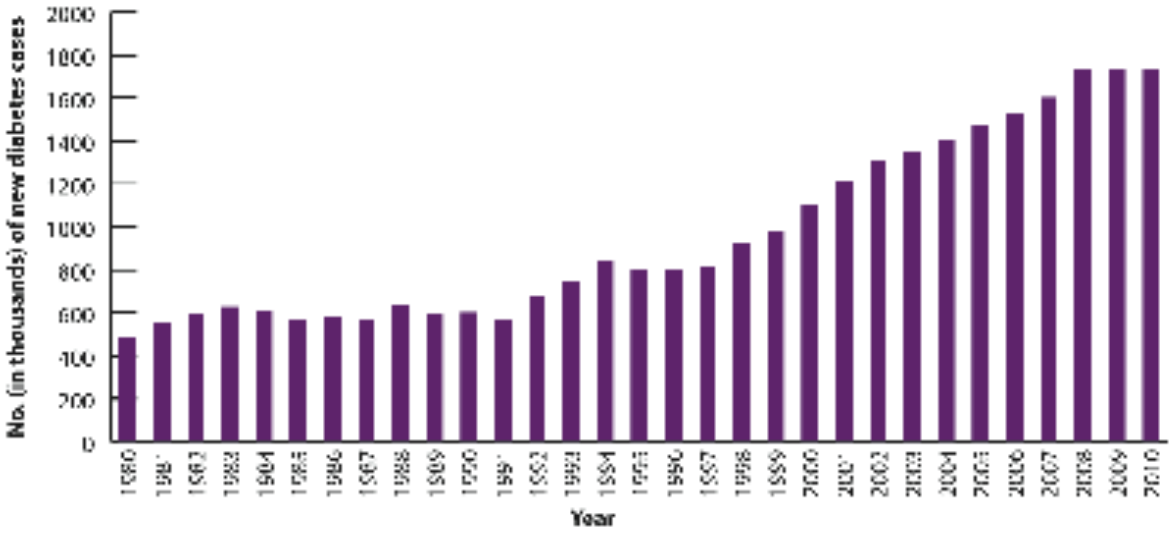

Figure 1-1 Annual Number of New Cases of Diagnosed Diabetes among U.S Adults Aged 18-79 [7]

1.2 Research Scope

In this thesis, a framework to analyze millions of tweets available from Twitter have been proposed to discover how socially shared food habit pattern of individuals are correlated with reported diabetes rates across the U.S. The main tool used for this purpose is sentiment analysis. Sentiment analysis is a classification technique that predicts positive or negative sentiments associated with posted tweets and typically uses Naïve Bayes or SVM as underlying classification method. In this work, a study have been done to discover how negative sentiments expressed in tweets containing the word “fast food” are correlated with reported diabetes rates across various states in the U.S. Note that, generic sentiment analysis technique uses a common

set of words, which are indicative of positive or negative sentiments, to predict the sentiment of a new tweet. The chosen hypothesis is that instead of using a generic sentiment analysis techniques, labeling the tweets containing the word “fast food” manually to train the classifier that performs sentiment analysis to observe the possibility of obtaining a better correlation. To test this hypothesis, this work uses both generic sentiment analysis as well as sentiment analysis engine trained on manually labeled training set to classify any new input.

The research performed as a part of this thesis can be broken down into the following steps. The first step is to collect tweets from the Twitter API. The next step is text preparation through Natural Language Processing (NLP). This step significantly affects the accuracy of analyses. The third step is to provide a proper training set to train the desired sentiment analysis classifier. Therefore aggregating a training statistics would needed in order to implement a proper learning procedure. The next step is to choose a proper classification method to apply on the corpus of data and then evaluate the predictive accuracy of a classifier. The final step is to compute the correlation coefficient between the predicted negative tweet sentiments and reported diabetes rates across various states of the U.S.

1.3 Thesis Organization

The rest of this thesis is organized as follows. In chapter 2, a brief introduction and statistics to Twitter is provided. In chapter 3, the sentiment analysis is described. Chapter 4 provides a literature review of related work. In Chapter 5, the proposed framework for the problem is described. Results are presented in Chapter 6 and finally Chapter 7 concludes the research and suggests future works.

CHAPTER 2

TWITTER, TWEETS AND DATA ANALYSIS USING TWITTER

In this chapter an overview of Twitter as a social media, tweets and twitter APIs have been developed. Also some of the common challenges of Twitter data analysis have been mentioned.

2.1 Twitter

Twitter contains a wide range of users from individuals to organizations and associations. According to the Twitter statistics, there are 302 million active users per month and 500 million Tweets per day. The majority of these users are located outside of the United States, let's said around (77%). Among these users 80% use their cell phone device to connect and use twitter.

The first step to access Twitter is establishing an account. Account holders could change some privacy on their own account, but unlike some social media such as Facebook, the default setting has been set as public. Each user has a personal profile page, which could be customized in different styles, for example account holders could change and customize some features such as the background, picture, web address, and brief biography. More than (~40%) of account holders just read messages and do not make posts, and those account holders who post could have multiple accounts [8]. After establishing an account, the account holder could post messages (tweets), repost messages (retweet), follow other account holders and attract other individuals to their account. Account holders could narrow down the criteria, based on which they are interested in streaming the related tweets and messages and they are even able to stream their interests, which are generally based on keyword searches. As soon as Tweeter figures out

the interest area of an account holder, it starts to algorithmically suggest relevant accounts and contents to follow [9].

One of the important feature of twitter is that all tweets have been limited to equal or less than 140 characters, therefore they would be referred to as micro blogs. Within the defied limitation on 140 characters, tweets could mention some URL links or other kind of links, webpages or blogs, pictures and hash tags. Tweets, which have not been defined for a specific targeted group of users, are called public tweets. Public tweets are available to any Twitter user and could also be sent to targeted user or group of users who are not followers. Under an account holder's policy and privacy, each tweet can be tagged to show its geographic origin. This significant feature could be combined with other with technological add-ons to help extracting sizable samples of valuable geo-located Twitter data [9].

Message distribution strongly depends on the number of followers who are linked to an account and the number of times that a message is retweeted. An account holder could see tweets in list format on his home page. Twitter timeline has been made by this chronological display [9]. From a research perception, text-based tweets should be considered data as well as the metadata that accompany them. Metadata is the data, which contains each account user's language, their geo-location, the number and names of the people they interested in, as well as the number and names of their followers [10].

Twitter uses two major kind of API, namely RSEST API and Streaming API. REST API, which offers programmatic access to read and write twitters data, and has two main functions to handle statuses: GET and POST. Streaming API, which provides low latency facility to access to Twitter's global stream of Tweet data. Twitter provides many streaming endpoints; each of which has been customized to certain use cases. User Stream, Site Stream and Public Stream,

streams of the public data flowing through Twitter. Public Stream is appropriate for particular users or subjects, and data mining including:

a. GET statuses / sample (1% of steaming data) Garden hose: Using Garden hose returns a minor accidental sample of all public statuses. The Tweets returned by the default access level are the same; therefore if two different users join to this endpoint, they will get the exact equal Tweets.

b. GET statuses / firehose: It just returns all public statuses. Just few researchers need this level of access. Productive combination of other methods and different access levels can satisfy approximately every request.

2.2 Twitter and Health Research Potential benefits and problems associated with Problems of Twitter data analysis

In health studies on twitter, researches have been divided into two domains, Big Data and little data [8]. Each domain has almost a different research method. This work focuses on health research area, which has been addressing through big data. One of the benefits of using twitter as a data set for data analysis is that twitter has the possibility to offer a massive amount of raw data, which makes the sampling error go to zero. Additionally when the sample size is huge, new methods and patterns are simpler and easier to identify, recognize and classify. Furthermore scholars are more interested in getting the correlations and probabilities compared to causal analyses so data mining from social media such as twitter could be the best option for such studies. Despite increasing attention in topic of big data and using twitter as a source of big data, large datasets such as Twitter and Facebook still are the essential part of challenges. The first issue is that data retrieved from Twitter are most likely of many unwanted errors. The reason that

Twitter contains this huge amount of errors is computer-generated spam that are located in Twitter as well as many texts and messages that contain lots of irrelevant information, unorthodox abbreviations and wrong spelling messages. Unfortunately, twitter methods for searching and retrieving data do not have 100% accuracy, for example for streaming tweets with combinations of filters, they have some errors, which causes a lot of problem to focused topic retrieval. To resolve such a major issue, self-correcting algorithms have been developed to get, clean and process data [8]. Another problem with the data that are collected from Twitter is that they would not be a definite good sample of general population. For instance, researchers have found that about half of Twitter users are adolescent, educated whose opinions do not necessarily shows the general public's thoughts about a special matter or topic. Furthermore, another problem is that governments and other powerful agencies can link. In spite of these drawbacks, Twitter still provides a very rich source of information, which when analyzed properly can help to discover many interesting patterns and associations.

CHAPTER 3

SENTIMENT ANALYSIS

Sentiment analysis is the main tool which is used in this thesis. In this chapter, a detailed description of sentiment analysis technique have been provided. Sentiment analysis, known as opinion mining, denotes the process of extracting subjective information from a source which contains objective and subjective information as well as other materials by applying NLP, text analysis and computational linguistics . Opinion mining is divided into three steps as follows. First, the input will be divided into two parts, afterward each part will be tested to realize if it covers any sentiment, meaning that each part will be examined to see if it is subjective or objective [11]. Second, the subjective sentences will be studied to distinguish their sentiment polarity. Finally, the objects of sentences that expressed an opinion might be extracted [12]. Opinion mining typically works on only positive and negative sentiment instead of working on discrete feelings and emotions such as happiness and sadness. Discrete emotions would not distinguish sentiment strength but they could help to enhance the accuracy of association of words with positive or negative sentiments [13]. Most of the opinion mining algorithms use machine learning techniques to classify general features related with positive and negative sentiment where the features could be a subsection of the words in the document, part of speech and so on [14] [15] [16]. Two machine learning challenges for sentiment are feature selections and classification algorithm choices. Feature selection is defined as processing data to remove the least useful n-grams in order to improve the accuracy of classifications, which will be described further. Sentiment analysis also could be described as a method for Natural Language Processing. NLP or NLU, Natural Language Understanding, is the subset of computational

linguistics, which itself is a combination of linguistics and computer science. In study of sentiment analysis, familiarity with some linguistic term can greatly help. Here, a few of such terms based on [17] have been defined

Lexicon: A set of defined words in a defined language is called lexicon. Each word could be classified through a lexical group or part of speech (POS) such as article, noun, verb, adjective, adverb, conjunction, preposition, or pronoun [17].

Syntax: Syntax handles the construction of a sentence from words. According to Chomsky's theory, syntax could be defined as terms of grammars and rules and it would not be dependent of semantic. However, syntax could also be explained by using some functions and words, which are related to each other, called dependency grammars. Dependency grammars greatly demonstrate efficient to parse texts and postulate a theoretical framework to many present parsing techniques, including many of which used for sentiment analysis [17].

Semantic: Semantic has been located exactly in opposite of syntax; means there are some complete sentences which are syntactically correct but that cannot make sense. For example: Colorless green ideas sleep furiously [17]. Semantic could change sentence sentiment greatly and semantic analysis can lead to a better sentiment analysis.

There are more realms of linguistics such as pragmatics, discourse, and dialogue; however, they have a slight effect on sentiment analysis of short informal sentences such as tweets.

During NLP tasks, which also include sentiment analysis, one of the hardest parts is handling ambiguity. Ambiguity is a problem, which could happen during all parts of linguistics. Some of the tools that could help clear ambiguity are probabilistic models, which are very

common in machine learning. The grouping of logic and statistical methods now enables us to parse running-text sentences with a success rate of nearly 90% [17].

For extracting subjective information from large datasets computational studies have been done about how opinions, attitudes, reactions, and perspectives are shown in different languages. It could provide important feedbacks for service providers, for example it could inform them of the feedbacks of a new product or some other criteria. Recent innovations even go beyond of measuring positive vs. negative, isolating a fuller spectrum of feelings and estimations and controlling for different subjects and community norms [18]. Because of exponential growing of social media and the point that many of our decision making processes are naturally social processes sentiment analysis become even more significant in recent (for example what you choose to eat or where you decide to eat are greatly depend on what your friends prefer or suggest.

It is clear that sentiment analysis is greatly useful but it has its own challenges. Sentiment focuses on subjective information and it depends on the manner and attitude of speaker on a special matter of subject and hearer inferences about that information – same as other kind of communication. Author and reader stances frequently contain very different but correlated sentiment information. Sentiment is blended and multidimensional. Sentiment is difficult and complex is both aspects of linguistically and socially. Conclusively, sentiment is context dependent [18]. These make sentiment analysis a tough task to accomplish.

There are few procedures and models required in order to accomplish sentiment analysis. First phase of sentiment analysis is text preparation. In text preparation, text will be prepared prior through one or more processes, such as tokenizing, stemming, negation and POS tagging in order to get it ready for sentiment analysis [18]. One of the important steps in text preparation is

tokenizing which is essential for sentiment analysis [18]. Researches could prove that using a suitable and accurate tokenizing can deeply advance accuracy of sentiment analysis. Another method for text preperation is called stemming . Stemming is a technique for collapsing distinct word forms. It could make a beneficial reduction in vocabulary size which is called features dimension [18]. Pang [19] explains that stemming would not be useful for reaching better accuracy in sentiment analysis. Next technique for text preperation is called negation. Sentiment words work very differently under the semantic range of negation (e.g. enjoy in ‘I enjoy it.’ vs. ‘I didn't enjoy it’) [18] .

Potts [18] has indicated handeling negation poperly is helpful for improving appropriate and more accurate sentiment analysis.

The policies of thumb for in what way negation relates with sentiment words are commonly as follows:

- Weak (mild) words such as good and bad has been considered like their opposites after negated: bad ≈ not good; good ≈ not bad
- Strong (intense) words like excellent and terrible have very wide-ranging meanings under negation: not excellent is constant with all from terrible to just-shy-of-superb, and different lexical things favor distinctive senses.

Negation’s expression is lexically different and its effects are far-reaching important challenge in task of negation [18]. Therefore deriving a general algorithm, which could handle negation, would not be an easy task [18]. And many surveyed researches do not focus on sentiment analysis of twitter

There are many cases in which there is a sentiment contrast between words, which have the same string representation then different parts of speech. This proposes could prove the value of implementing a part-of-speech tagger on your sentiment documents and afterward applies the resulting word–tag pairs as features or factors of features [18]. Though it is hard to find a clear conclusion in surveyed papers on effectiveness of POS-tagging in sentiment analysis of tweets.

Next step of sentiment analysis after text preparation would be classifying texts based on their sentiment. On this phase, a sentiment lexicon and a classifier model would be necessary to accomplish the task. A list of words and their respective sentiment (good or bad) for a classifier could be provided by a sentiment lexicon. There are many unrestricted and exclusive sentiment lexicons for many different subjects like food, are available. Furthermore, many approaches are available for creating sentiment lexicon from scratch (please refer to [18].

Classifier models perform some tasks to classify texts to various sentiment classes (typically positive or negative). Mainly, a sentiment analysis classifier would take a text as an input, perform some process on it, and categorize it as positive or negative and preset it as its output. There are many types of classifier. The most practical ones in domain of sentiment analysis would be Naive Bayes (and its variations), Maximum Entropy (MaxEn), and Support Vector Machines (SVM). Naive Bayes classifier is the cleanest and simplest trained, probabilistic classifier model. In many of surveyed papers, they could have proved that using Naïve Bayes has led into suitable results for sentiment analysis of twitter. A simple algorithm for training a Naive Bayes classifier by just the words as features would be:

- Approximation of the probability $P(c)$ of every class $c \in C$ by dividing the number of arguments or words in documents in c by the whole number of arguments in the corpus.

- Approximation the likelihood distribution *P(w | c)* for all arguments or words *w* and classes *c*. This could be completed by dividing the number of tokens of *w* in documents in *c* by the whole number of words in *c*.
- To grade a document *d* for class *c*, approximation:

$$score(d,c) \stackrel{\text{def}}{=} p(c) * \prod_{i=1}^{n} p(w_i \,|c) \qquad 1)$$

- To calculate the most probable class label, pick the *c* with the maximum score value. To reach a likelihood of distribution, compute:

$$p(c|d) \stackrel{\text{def}}{=} \frac{score(d,c)}{\sum_{c' \in C} score(d,c')} \text{ [18]} \qquad 2)$$

The "naive" assumption is supposing each character to be independent of all other characters. This is a suitable assumption in classifying short texts (such as tweets) using bag-of-words [18].

Naïve Bayes would be a supervised classifier, which means that it must be trained and then applied. Creating a training set to lead into classifier could be a challenging task (which has been addressed in implementation part).

Conclusively, measuring effectiveness of a classifier would be necessary. Therefore, series of measures have been developed each of which has it is own advantages and disadvantages. To list a few:

- Accuracy: correct guesses divided by all guesses
- Precision: accurate guesses penalized by the amount of inaccurate guesses
- Recall: correct guesses penalized by the number of missed items

In overall, no single measure is good enough to be used on its own for evaluating classifier effectiveness (since they can be cheated). Following one can find how to cheat abovementioned measures:

- Accuracy: If the classes are highly imbalanced (case with most real world data) one can get high accuracy by often predicting the biggest group.
- Precision: One could frequently get high accuracy for a group *C* by infrequently predicting *C*, but this will ruin your recall.
- Recall: One could frequently get high recall for a group *C* by constantly predicting *C*, but this would ruin your accuracy.

So, derived measurements, such as:

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)} \quad 3)$$

CHAPTER 4

LITERATURE REVIEW

This chapter provides details literature reviews that are related to the work done in this thesis. Two types of literature reviews which are related to, (a) use of social media for public health research and (b) sentiment analysis have been provided.

4.1 Public health research and Twitter

Social media and Twitter have been used to address many public health research problems. For example, in [20] the authors apply a kind of Ailment Topic Aspect Model to over one and a half million health related tweets in order to discover mentions of over a dozen ailments such as allergies, obesity. Chew and Eysenbach (2010) in [21], by data mining from Twitter monitor public view of the 2009 H1N1 pandemic in order to monitor the use of the terms “H1N1” versus “swine flu”, sentiment analysis of “tweets”; and use Twitter as a real-time data resource, sentiment, and public attention trend-tracking tool.

In [22], the authors study about Twitter status updates regarding to "antibiotic(s)" to study and explore evidence of misunderstanding or misappropriation using of antibiotics. In [3], the authors predict the flu trends and in [4], the authors predict age of smoking by data mining from social media. In [5], the authors identify the helpfulness of Facebook and Twitter in relation to groups with Colorectal Cancer (CRC), Breast Cancer and Diabetes Mellitus. In [6], Eichstaedt et al. used language expressed in twitter to explain the “community-level psychological correlates of age-adjusted mortality from atherosclerotic heart disease (AHD) ". They identified that the language pattern showing negative different moods are risk factors for AHD whereas as positive emotions could be good factors against AHD.

4.2 Sentiment Analysis

Over the past few years, a lot of work has been done in the area of sentiment analysis. In [23], a novel method, which no human effort is needed to classify the documents, has been introduced to collect a corpus with objective texts and a corpus with positive and negative sentiments. By using Twitter API a corpus of text has been collected to make a dataset of three groups: positive sentiment, negative sentiments and objective texts but no sentiments. Afterward, a query has been provided for two types of emotions: first, happy emotions which includes " :-), :), =), :D" etc. Second, sad emotions such as " :-(, :(, =(,;(" etc. These two kinds of collected corpora were used to train a classifier to recognize positive and negative sentiments. Moreover, as each comment could not be more than 140 characters on tweeter it is considered as single sentence. So any emotion that has been found on a message is assumed an emotion for the whole comment. Afterward some statistical methods has been applied to linguistic analysis of the collected data, which has been gotten from Twitter API, to make a sentiment classification method for micro blogging [23].

In [24], a couple of web-blogs has been used to make corpora for sentiment analysis. The authors used emotion icons that have been allocated to blog comments to determine the users' emotion and mood for a specific subject. Afterwards by applying SVM and CRF learners they could implement a classifier at the sentence level and then examined different methods and strategies to reach to the final sentiment for a document. As a conclusion, considering the sentiment of the last sentence of a document as the sentiment at the document level could show the best result.

In [25], the authors used emotion faces such as ":-) for positive and :-(for negative " to make a training set for sentiment classification. A group of texts which containing emotions has

been collected from newsgroup and then the data set has been divided to two groups: positive and negative. Afterward SVM and Naïve Bayes showed 70% on the test set.

In [26], the approach of emotion based to implement sentiment analysis on Twitter and they performed different classifiers have been used. And at the end the authors found Naïve Bayes classifier could give the best results. It came up to 81% of accuracy on the dataset that has been tested. However they could not get a good result when they used three classes “positive, negative and neutral “

In [26], a classification of tweets have been done by using the approach of happy and sad emotions. And for feature space unigram and bigram method have been applied. Afterward the authors proved that unigram model overtakes the bigram approach. They biased their collected data by applying search queries in Twitter also they applied POS (part of speech) tags as features. Finally, they proved that applying POS tags would not be useful in classification.

Barbosa et al in [27] introduced a novel approach in sentiment analysis classification. Polarity predictions as noisy labels from three different websites have been applied in order to train a model. The desired data set have been divided to two parts, one part has been contained 1000 manually labeled tweets for modification and the second part has been contained 1000 manually labeled tweet for testing. Next syntax features of tweets with polarity of words and POS of words have been applied. As more abstract representation of tweets instead of raw messages have been used the authors could get a higher accuracy in sentiment classification.

In [28], a sentiment analysis on Twitter data have been done. By using real valued prior polarity and combing prior polarity with POS they extend the approach of [27]. Also in [28], a sentiment analysis on feedback data from Global Support Services survey have been applied in

order to get the effect of linguistic features such as POS tags. Extensive feature analysis and feature selection have been applied. The results demonstrate that abstract linguistic analysis features contribute to the classifier accuracy. In [28], a broad feature analysis have been performed. The results illustrate that the applying just 100 abstract linguistic features in each form could get the same outcome as using hard unigram based line. So the results proved that combing prior polarity could enhance the accuracy of their classifier, and concluded that tweet syntax features might help but only slightly.

CHAPTER 5

SENTIMENT ANALYSIS FRAMEWORK

In this chapter a detailed overview of the research have been provided. As mentioned before in the introduction, the research performed as a part of this thesis can be broken down into the following steps. The first step is to collecting tweets from the Twitter API. The next step is text preparation through Natural Language Processing (NLP). This step significantly affects the accuracy of analyses. The third step is to provide a proper training set to train the desired sentiment analysis classifier. Therefore aggregating a training statistics would needed in order to implement a proper learning procedure. The next step is to choose a proper classification method to apply on the corpus of data and then evaluate the predictive accuracy of a classifier. The final step is to compute the correlation coefficient between the predicted negative tweet sentiments and reported diabetes rates across various states of the U.S. a description for each of these steps below has been provided.

5.1 Streaming Data from Tweeter Using REST API

In the very first phase more than 130,000 tweets which contain the word “Junk Food”, have been streamed through the REST API. “Junk Food” is a pejorative term for food of little "nutritional value" and excessive fat, sugar, salt and calories. Junk food might also refer to high protein food containing large amounts of meat prepared with, for example, too much unhealthy saturated fat. Therefore, no sentiment analysis is performed to separate the negative, positive and neutral tweets in this phase and we consider all tweets show negative semantic. Location is identified through user profile information, which has been given by users and might not be

completely accurate for example in cases they use some acronym words, miss information or wrong dictation.

5.2 Streaming Data from Tweeter Using Stream API

In the second phase, instead of using the previous method for streaming, extracting and interpreting data is improved by developing a framework to implement a real-time sentiment analysis on stream of tweets based on Streaming API.

5.3 Data Processing for Streaming Tweets

In order to receive more useful tweets this approach starts streaming tweets by filtering streams based on a given language, which in this work is English and a given geographical bounding box coordinates which can be used to approximately limit origination point of tweets to a specific region, which in this work is the US. Then connecting to “Map Quest” API, the state that the tweet has been originated from, and including a given key word, which was “fast food” is extracted.

Changing the key word from “Junk food” to “Fast food” is done in order to implement better sentiment for every single streamed tweet.

5.3.1 Reordering Streamed Tweets

After applying the desired filters, more than 100,000 tweets in 24 hours of streaming are collected by using Stream techniques to collect data from tweeter, which is a better and more practical methods for studying on big corpus of data in order to collecting more amount of data in a faster way. Python 2.7 and Twitter API are used for streaming, cleaning and processing the desired data.

5.3.2 Switching From REST to Stream API

Switching from REST API to Streaming API is done in order to do real time streaming in a fast way. When there is a long-lived HTTP connection, the Streaming API constantly sends new responses to REST API queries. Moreover, the Streaming API gets updates on the latest Tweets and syncs itself with user profile revises and more. Therefore when the application is rate-limited for over-polling the REST APIs the Streaming APIs would be better and definitely faster option for streaming tweets in a real time manner [29].

5.3.3 Extracting the Exact Location of Tweets by MapQuest API

In the next phase the exact state which every single tweet is originated from is identified using Map Quest API. MapQuest retrieves the boundary box of each tweet, finds the state of that tweet and then in case of success, exports it to a CSV file for next processes.

5.4 Proper Classification Method for Tweets

Next challenge of sentiment analysis of twitter data is choosing a proper classification method. Studies express that most classifier models, such as Naïve Bayes and MaxEnt, has an acceptable performance for positive/negative classification of tweets. However, neutral/sentimental classification can be challenging. Here at this sentiment analysis of tweets should be implemented to see either the attitude of the tweet about “Fast Food” is positive, negative or neutral. This is a big challenge to do a good sentiment for a specific topic, as there are many mis-information and many fake reviews between the tweets that people put on their twitter accounts. Preparing tweets for feeding into classifier regardless of it being for training, testing, or classifying is a significant task. Many considerations should be applied to the tweet before leading it to classifier. The first step for starting sentiment analysis is choosing a classifier

that could classify each tweet according to its attitude. In this work each tweet is assigned to one of the three categories, positive, negative and neutral.

5.4.1 Classification Methods

In order to gain more accurate results, two different methods for classification and sentiment on the same dataset are performed. Fist, using the simple sentiment done by "TextBlob" package and second, providing manual trained classifier and labeled test set.

"*TextBlob"* is a Python library which is used for processing and handling the data that works with human language or textual data by providing an API that could dive into natural language processing (NLP) tasks [1]. The sentiment property calculates the polarity score for each tweet which would be a float number within the range [-1.0, 1.0]. The subjectivity would be a float number within the scope [0.0, 1.0] where 0.0 is definite objective and 1.0 is definite subjective [1]. Also TextBlob Classifiers module makes it simple to create custom classifiers. As a method of classifying many papers have been surveyed that explained in literature section and conclude that the best method between supervised classification methods for big corpus of data is Naïve Bayes Classifier. Thus the streamed tweets are passed through the TextBlob Naïve Bayes classifier and categorized for 14 chosen states in the US.

5.4.1.1 Generic Classifier

In this work the generic classifier of TextBlob without providing any training set that requires manual labeling have been used.

5.4.1.2 Classifier using manually labeled training set

As another method of classification, a manually labeled training set for training the classifier have been provided. A training set containing 3000 tweets (each containing the word "Fast Food") have been built and their sentiments have been manually labeled. This training set

was used to train the sentiment analysis classifier. The predicted sentiments obtained on the test set was then categorized according to various states obtained by using geo-tag location.

5.5 Evaluating Classifiers

Final challenge is to evaluate the predictive accuracy of a classifier. Using data mining for streams, the best regularly used degree for calculating analytical accuracy of a categorizer is prudential exactness [30]. It is discussed that this degree would be just suitable if all classes are stable, and will offer (approximately) the similar number of instances [31]. This means that the level that Streaming API at twitter transports positive or negative tweets through might differ by time. It is not possible to suppose that it is constantly 50%. Though, a degree that by default takes care of fluctuations in the class dissemination needs to be desirable [31]. Thus it is needed to provide a manually labeled test set to pass it through our classifier in order to evaluate the accuracy rate of both the manually trained classifier and the automatic classifier. The predictions can be compared to the class labels in the test dataset. Accuracy of a classifier could be a ratio between 0 and 100 % [32]. 2000 tweets as a test set have been trained, and they have been passed through the classifiers. The automatic generic classifier gives the accuracy rate of 29% and the manual classifier gives the accuracy rate of 60% while 3000 labeled tweet has been used as the training set and 2000 labeled tweets has been used as the test set.

5.6 Matplotlib Package

The next step is to choose a library to plot different graphs and sketches to see the final results and correlations. Thus, Python 2.7 and its popular package for graphical works, “Matplotlib” is used. Matplotlib is a python 2D plotting library that helps making figures in a

variety of platforms, which could be used in python scripts, ipython shells, web application servers, and six graphical user interface toolkits [33].

CHAPTER 6

RESULTS

In this chapter an experimental evaluations of the proposed framework using both generic and manually labeled classifiers have been provided. Then a report of correlation coefficients between negative sentiment expressed in tweets containing the word “fast food” and reported diabetes rate for 14 different states of the United States have been proposed. The reason for restricting the analysis to these 14 states was that among the collected tweets across various states of the U.S., these 14 states had sufficient number of tweets containing the word “fast food”. For the experiments 100000 streamed tweets have been used. Geotagging have been used to extract the USA States of from these 100,000-streamed tweets which previously have been filtered by location, USA, language, English and keyword, fast food. Next, two types of classifiers have been trained, generic and manually labeled and they have been used to predict state-wise positive, negative or neutral sentiment percentages from these 100,000 streamed. Next, correlation coefficients between negative sentiments and reported diabetes rates for these 14 states have been computed. As the results will be presented next, manually labeled classifier shows stronger correlation among negative sentiments and diabetes rates as compared to generic classifier.

In section 6.1, the results using generic classifier have been provided and in section 6.2, the results using manually labeled classifier have been provided.

6.1 Results for Generic Classification with TextBlob:

Figure 6.1 shows the percentage of Positive, Negative and Neutral tweets among 14 US states when the tweets have been passed through generic TextBlob classifier. It clearly shows

that Alabama has the least percentage of Negative attitude about fast food and Washington and Oregon has the most negative attitude about fast-food

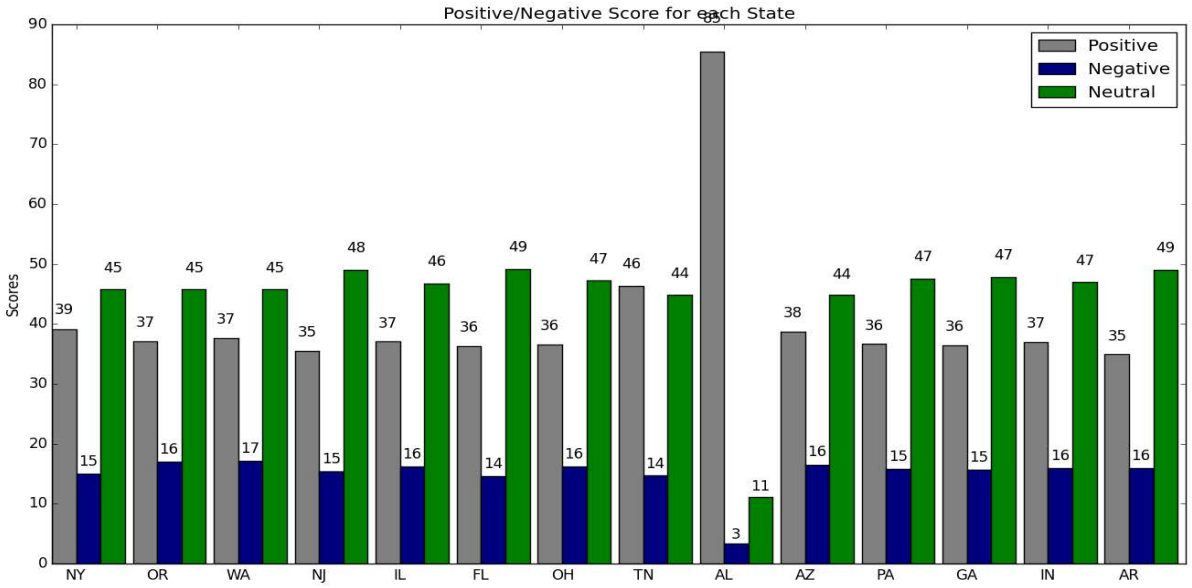

Figure 6.1. Sentiment result for 14 states

Figure 6.2 shows the negative percentage attitude about fast-food and the percentage of adults who have been diagnosed with diabetics in US in 2012 according to [7].

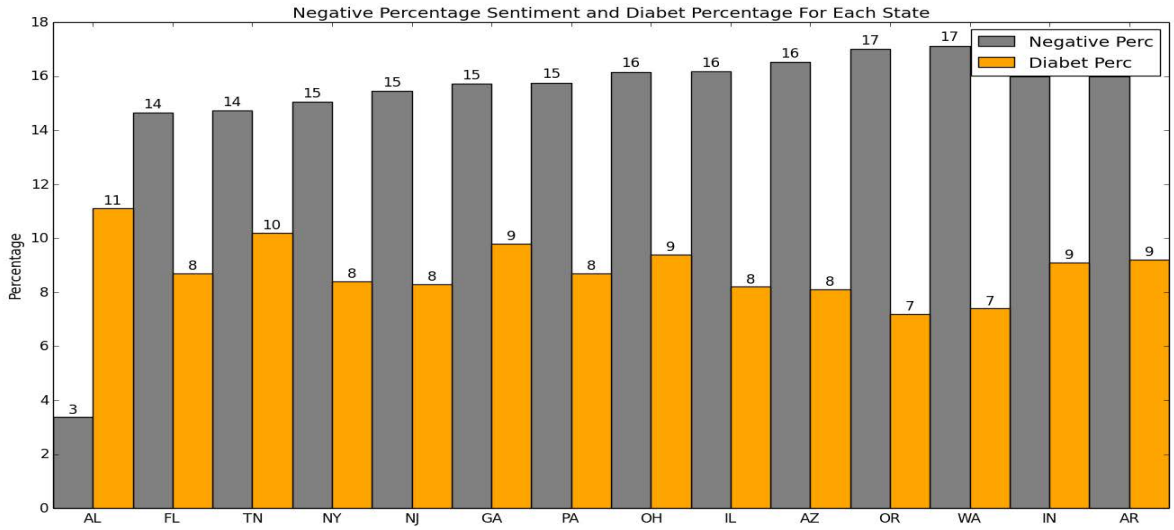

Figure 6.2. Comparison between Negative Sentiment and Adults with Diabetes

Figure 6.3 shows the portion of negative attitude in each state. Oregon and Washington have the most portions of the whole negative tweets about fast food among other 14 US states.

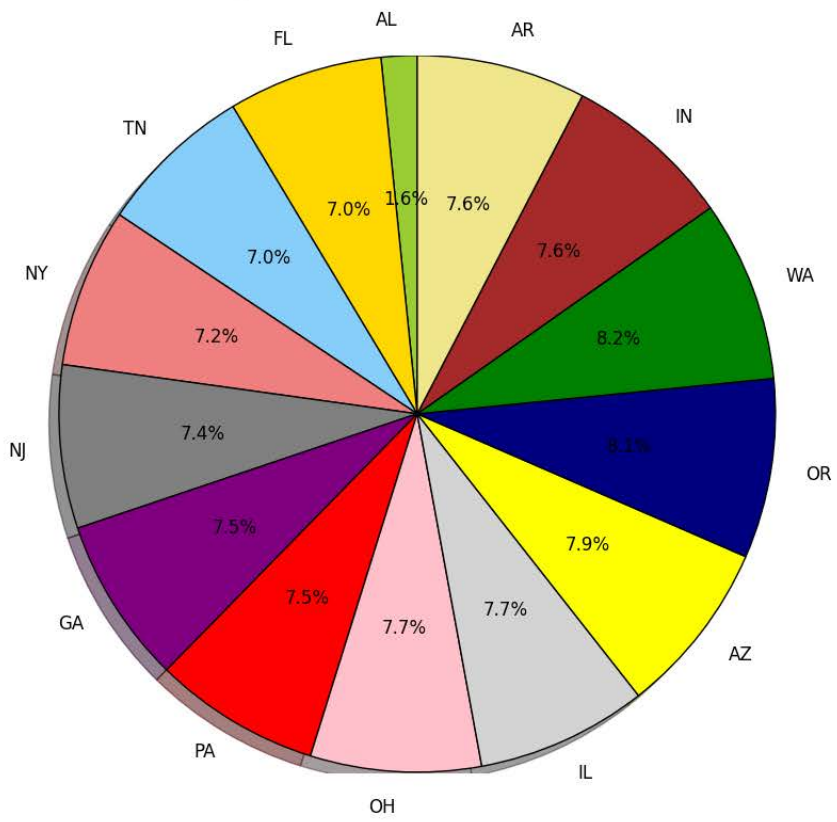

Fig 6.3. Negative attitude portion in each state

Figure 6.4 shows the percentage of the negative attitude about fast food for chosen states and the statistical percentage of adults in those states diagnosed with diabetes. This plot shows a descending trend, meaning when the percentage of negative attitude in a state increases, the percentage of adults with diabetes in those states decreases.

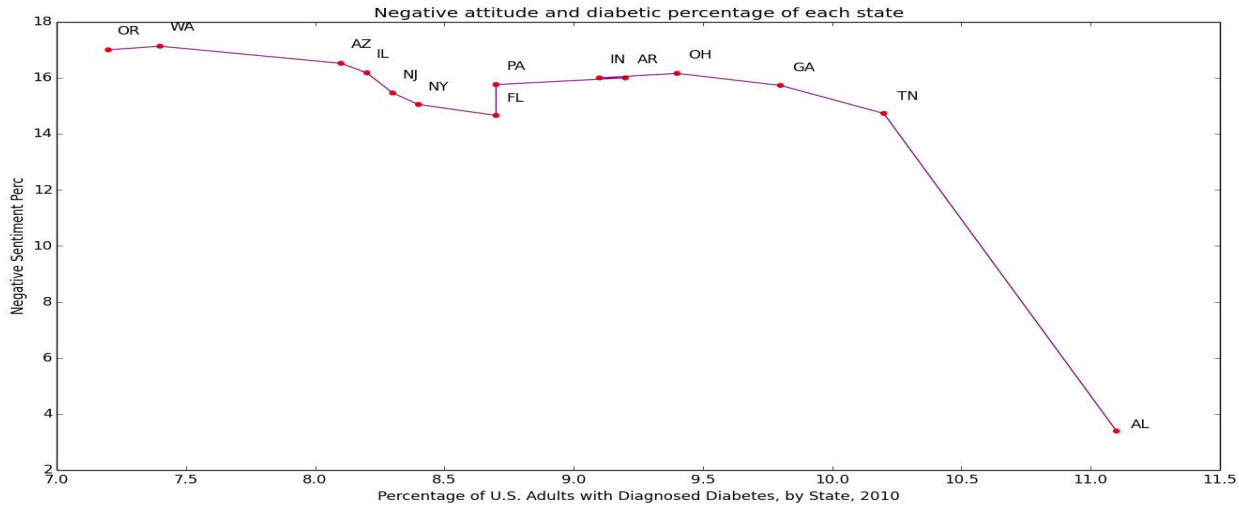

Figure 6.4. Scatter plot for negative attitude and diabetes percentage for each state

Figure 6.5 shows the linear regression model results (-66% correlation coefficient) for TextBlob generic classifier for chosen states according to their percentage of negative attitude about fast food. A linear regression line has an equation of the form 4 where ***X*** is the explanatory variable and ***Y*** is the dependent variable. The slope of the line is ***b***, and ***a*** is the intercept [34].

$$Y = a + bX \qquad (4)$$

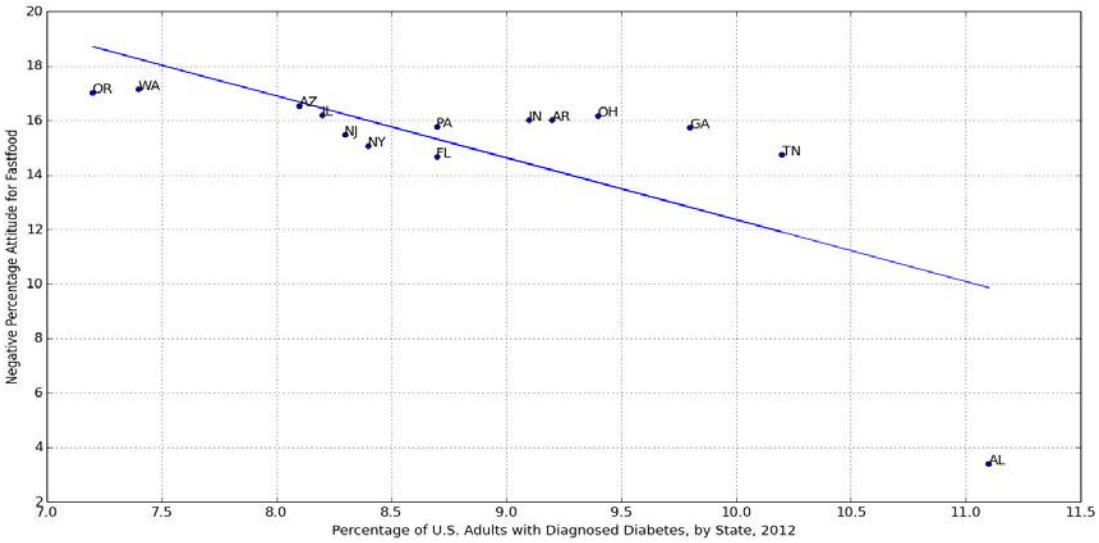

Fig. 6.5. Linear regression model for predicting diabetes rate

To fit the data between negative sentiments and diabetes rates for these 14 states, an order 4 polynomial regression that captures non-linear dependence have been used. Polynomial regression is a kind of linear regression which tries to model a non-linear relation between the independent variable *x* and dependent variable *y*. It provides a model to show a nonlinear phenomenon by donating a nonlinear relation between the value of x and the corresponding conditional mean of y [35]. Polynomial regression models usually used least squares method to minimize the unbiased estimators' variance of the coefficients. Sometimes data fits better with a polynomial curve. In this work a quadratic model has been applied in order to show a better trend of our data, although it is straightforward to extend this to any higher order polynomial.

Figure 6.6 shows the trend of correlation between negative attitude about fast food and diabetes percentage**.**

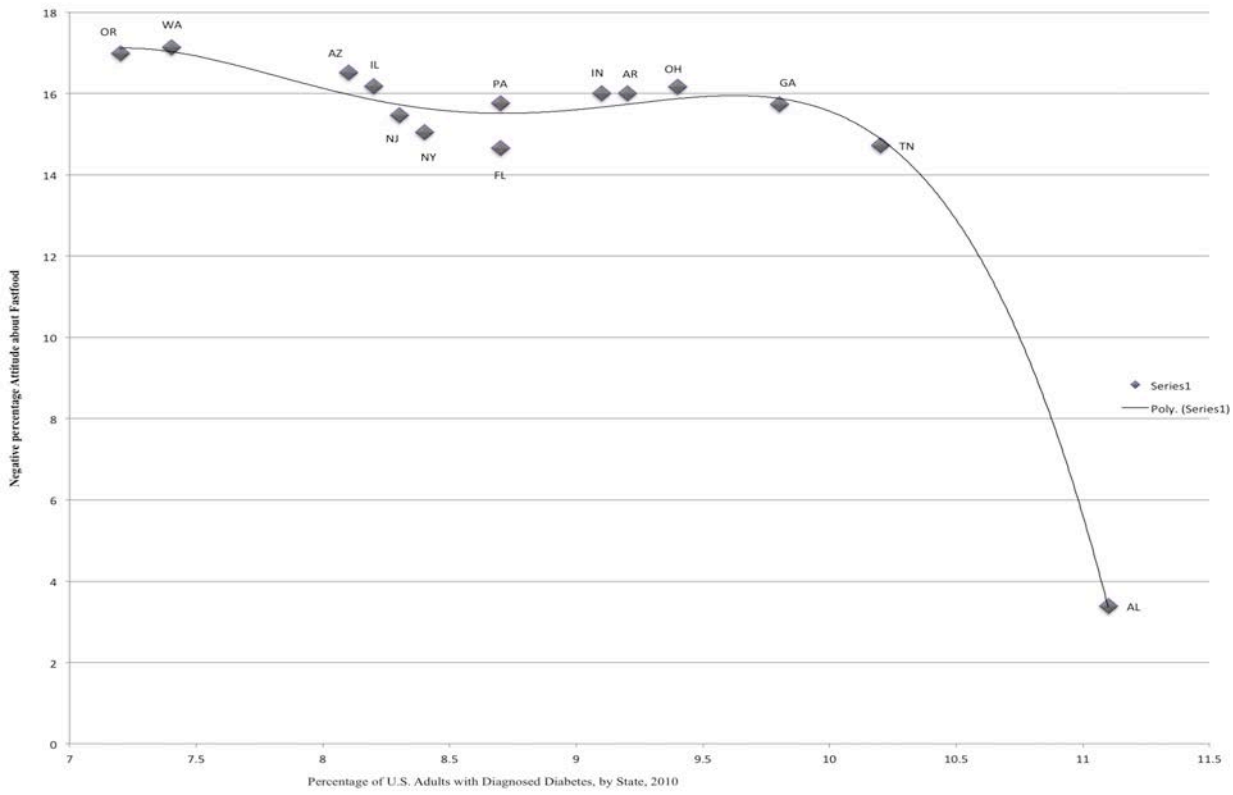

Figure 6.6. Negative attitude about fast food for 14 States

6.2 Results for Manually labeled Classifier with TextBlob:

Figure 6.7 shows the percentage of Negative and Neutral tweets among 14 US states when the tweets have been passed through manually labeled classifier TextBlob classifier. For manually trained classifier a training set of 3000 tweets have been used, among these tweets there are many advertisements and comments regarding to thousands fast food restaurants such as, “tacobell becomes first fast food chain with vegetarian-certified” or “Breakfast: Fast and go!Yayyy fastfood McDonald” and obviously they are positive sentiment, moreover our focus in this work is on Negative attitude regarding fast food ,therefore we combined positive and neutral percentage and consider them as Neutral. Again manually labeling shows that Alabama again has the least percentage of Negative attitudes about fast food and Washington and Oregon has the most negative attitude about fast food

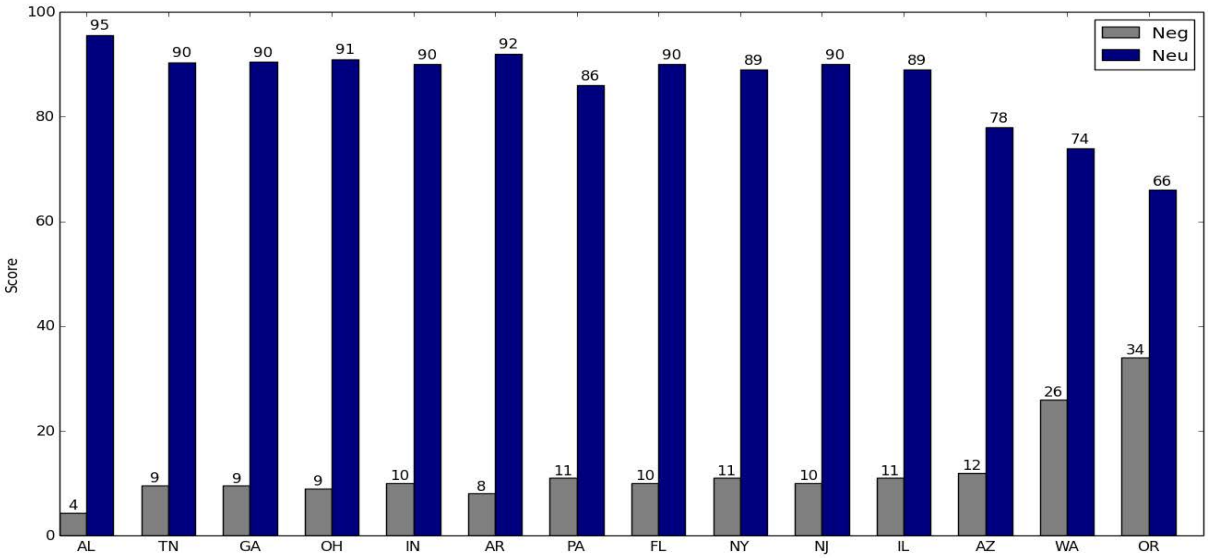

Figure 6.7 Sentiment Result For 14 States

Figure 6.8 demonstrates the portion of negative attitude between 14 states for manually trained classifier. So similar to the simple sentiment method results, this graph shows that

Oregon and Washington have the most portions of the whole negative tweets about fast food among other 14 states in the US.

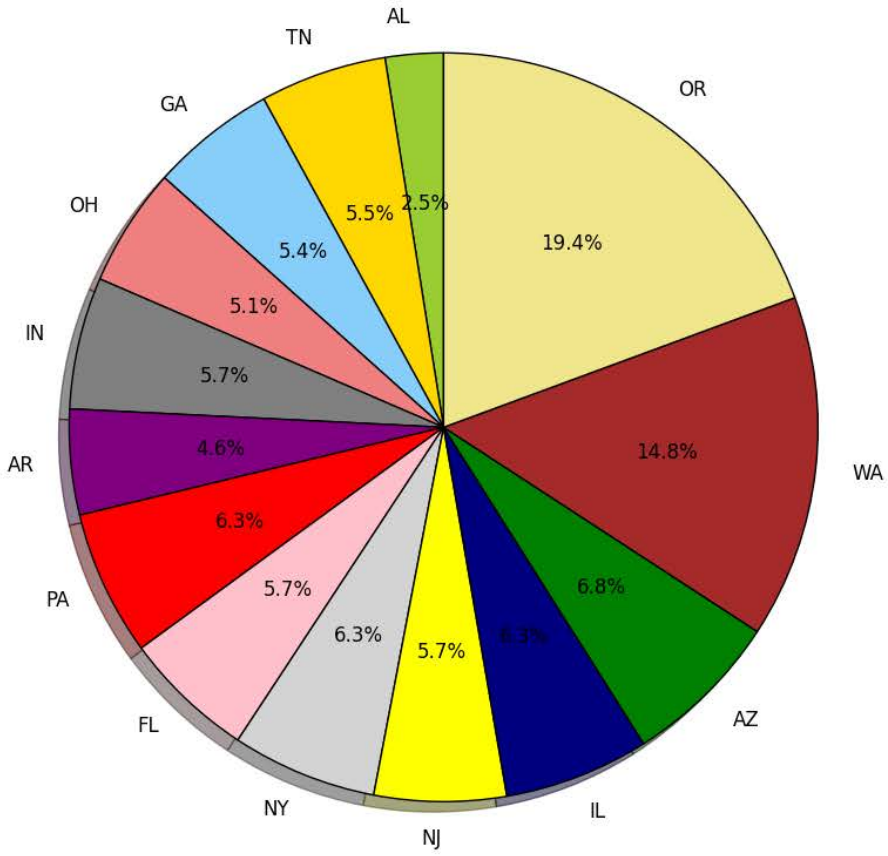

Figure 6.8. Negative attitude portion in each state

Figure 6.9 shows the negative percentage attitude about fast-food and the percentage of adults who have been diagnosed with diabetics in US in 2012 according to [7].

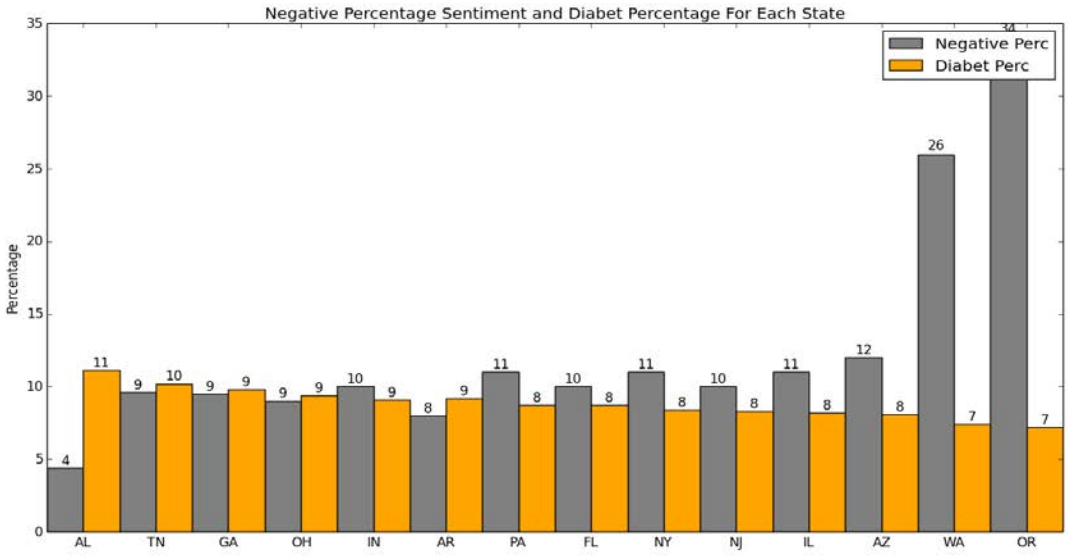

Figure 6.9. Comparison between Negative Sentiment and Adults with Diabetes

Figure 6.10 shows the percentage of negative attitude about fast food for chosen states and the statistical percentage of adults in those states who have been diagnosed with diabetes. Here manually trained TextBlob classifier have been used while in figure 4 automatically trained TextBlob classifier have been used. This plot has an obvious descending trend, stating that when the percentage of negative attitude in a state increases, the percentage of adults with diabetes decreases for those chosen states.

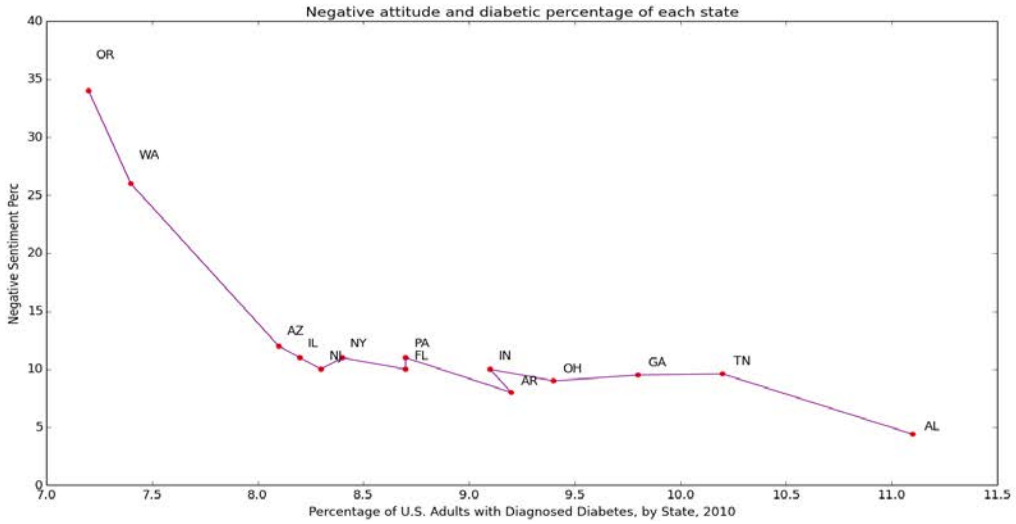

Figure 6.10. Scatter plot among 14 states for manually trained model

Figure 6.5 shows the linear regression model results of generic sentiment for chosen states according to their percentage of negative attitude about fast food. Here manually trained classifier is used while in figure 6.5 generic classifier. It shows (-72%) correlation which is significantly greater than the correlation coefficient in generic classification TextBlob method.

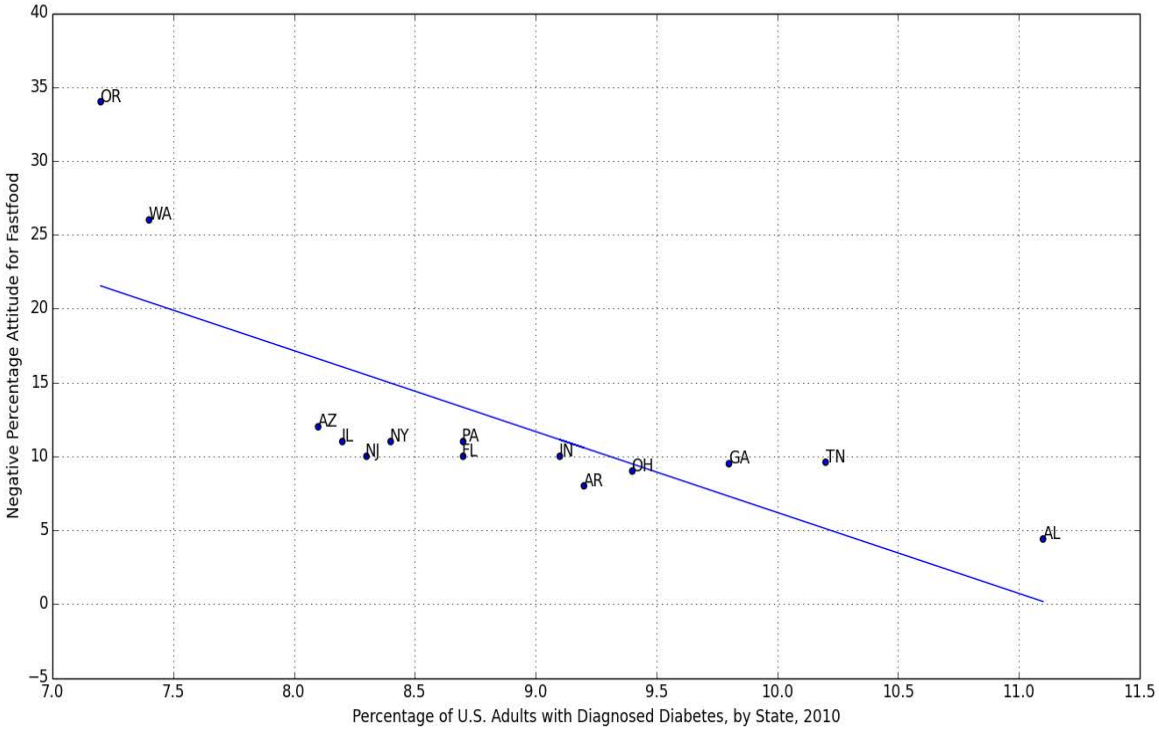

Figure 6.11. Linear regression model for 14 states for Generic TextBlob

As linear regression could not be expected to show an exact linear dependency, the correlation between negative attitude about fast food and actual percentage of adults with diabetes have been shown by using the regression trend of order 4 in Figure 6.12. This diagram shows almost descending trend between these two variables.

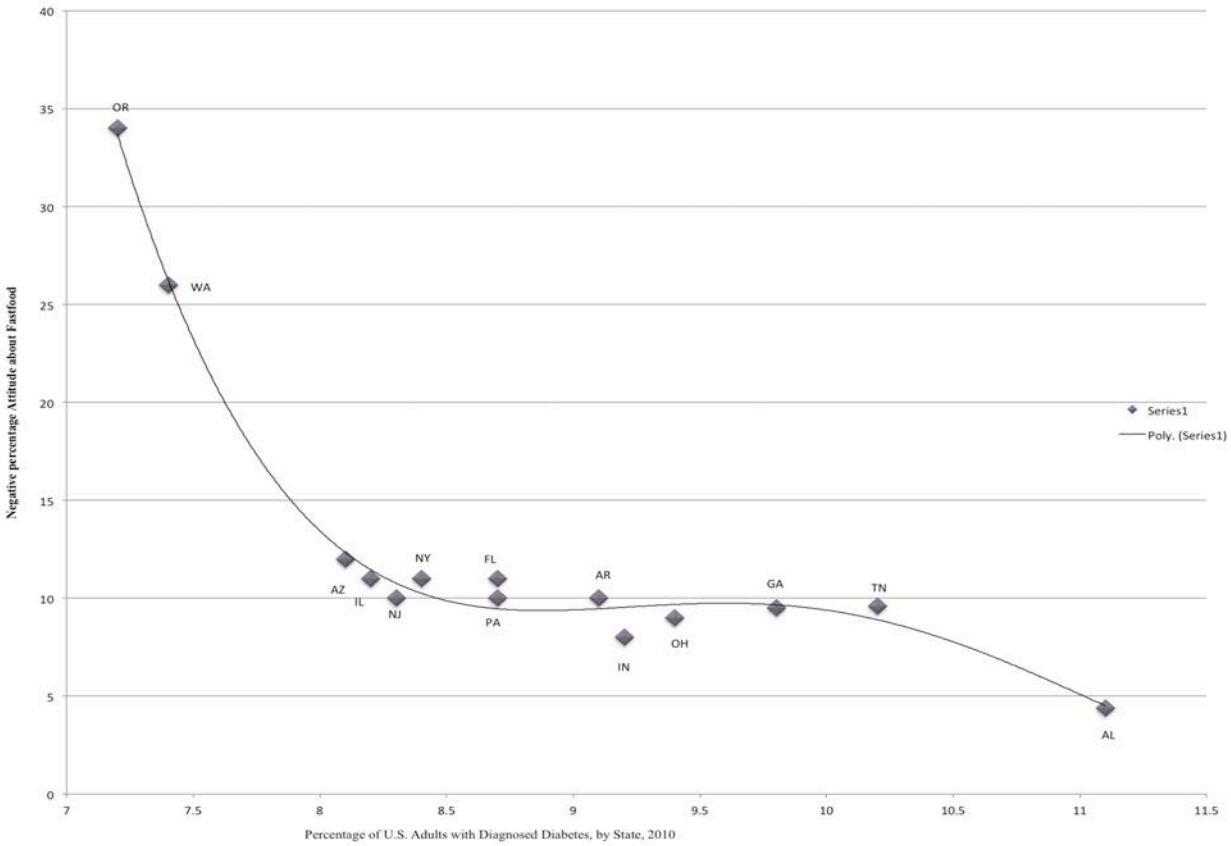

Figure 6.12. Regression trend of order 4 for manually trained model for 14 states

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this study, a framework which performs real-time sentiment analysis on streamed tweets, have been developed. Connecting to twitter through Streaming API, this framework starts receiving tweets by filtering streams based on a given language, English, a geographical bounding box coordinates, US, and a keyword, “Fast food”, in order to study the correlation between eating behavior of people in specific states and the percentage of people that has been diagnosed with diabetes in those state.

The experimental results show (29%) accuracy for generic classifier and (60%) for manually labeled classifier. For both the classifiers, correlation coefficients between predicted negative tweets percentage and reported diabetes rates were computed. It was observed that negative sentiments predicted by the manually labeled classifier showed stronger correlation (-72%) to the reported diabetes rate for 14 states of the U.S. as compared to that of the generic classifier showing (-66%) correlation. Therefore, the results support the hypothesis that manually labeled classifier will perform better as compared to generic classifier in terms of prediction accuracy and obtained correlation coefficient. This result was not surprising as training set for generic classifiers were not really associated with specific sentiments related to food habit patterns but were associated with general sentiment, whereas the training set for manually labeled classifiers were specifically labeled to be associated with food habit pattern.

Another important result was the descending trend of both methods for chosen states. Although both methods show descending trend between negative attitude of using fast food and actual rate of diabetes, the manually trained classifier method shows more reasonable trend

compared to generic sentiment method. Moreover, manually trained classifier showed (-72%) correlation coefficient while generic showed (-66%) correlation. So manually labeled classifier showed stronger correlation between negative sentiment expressed in tweets containing the word “fast food” and reported diabetes rates as compared to generic classifier

The future research could be extended in three possible directions. First, by making only a minor change, this framework can be used to study association between sentiment expressed in social media and various other public health issues and/or reported disease rates. Second, by performing the analysis on a much larger scale by collecting tweets from all the 50 states of the United States. Third, by providing a larger training containing more manually labeled tweets.

REFERENCES

REFERENCES

[1] Steven L. TextBlob. [Online]. https://textblob.readthedocs.org/en/dev/ . [cited Oct 20,2015]

[2] Asur S. and Huberman B.A., "Predicting the Future with Social Media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, Toronto, ON, Aug. 31 2010-Sept. 3 2010, pp. 492 - 499.

[3] Achrekar H., Gandhe A., Lazarus R., Ssu-Hsin Y., and Benyuan L., "Predicting Flu Trends using Twitter data," in *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, Shanghai, 2011, pp. 702 - 707.

[4] Ungera B. and Chena X., "The role of social networks and media receptivity in predicting age of smoking initiation: A proportional hazards model of risk and protective factors," in *a Institute for Health Promotion and Disease Prevention Research, University of Southern California School of Medicine*, vol. 24, Los Angles, 1999, pp. 371–381.

[5] De la Torre-Díez I., Díaz-Pernas F., and Antón-Rodríguez M., "A content analysis of chronic diseases social groups on Facebook and Twitter.," in *Telemed J E Health.*, vol. 18, Spain, 2012, pp. 404-408.

[6] Eichstaedt J.C. et al., "Psychological Language on Twitter Predicts County-Level Heart Disease Mortality," in *Psychological Science*, vol. 26, 2015, pp. 159-169.

[7] Division of Diabetes Translation, "Diabetes Report Card," National Center for Chronic Disease Prevention and Health Promotion Centers for Disease Control and Prevention, Atlanta, Report Card 2012.

[8] Finfgeld-Connett D., "Twitter and Health Science Research," *Western journal of nursing research*, vol. 37, pp. 1269-1283, 2015.

[9] Twitter Inc., support.twitter. [Online]. https://support.twitter.com/groups/50-welcome-to-twitter . [cited May 10, 2015]

[10] Twitter, Inc. Twitter. [Online]. https://about.twitter.com/company . [cited May 10, 2015]

[11] Pang B. and Lee L., "Sentimental education:Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceeding of the 43rd Annual Meeting of the ACL*, Stroudsburg, 2004, p. 221 of 619.

REFERENCES (continued)

[12] Gamon M., Aue A., Corston-Oliver S., and Ringger E., "Pulse:Mining customer opinion from free text(IDA 2005)," in *Lecture Note in Computer Science*, vol. 3646, Madrid, 2005, pp. 121–132.

[13] Kaji N. and Kitsuregawa M., "Building lexicon for sentiment analysis from massive collection of HTML documents," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, 2007, pp. 1075--1083.

[14] Abbasi A., Chen H., Thoms S., and Fu T., "Affect analysis of Web forums and Blogs using correlation ensembles," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, 2008, pp. 1168-1180.

[15] Ng V., Dasgupta S., and Arifin S.M.N, "Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews.," in *Proceedings of the COLING/ACL 2006 Main Conference*, Stroudsburg, 2006, pp. 611-618.

[16] Tang H., Tan S., and Cheng X., "A survey on sentiment detection of reviews," in *Expert Systems with Applications: An International Journal*, vol. 36, 2009, pp. 10760-10773.

[17] Nugues P.M., "An Overview of Language Processing," in *Language Processing with Perl and Prolog*.: Springer-Verlag Berlin Heidelberg, 2014, pp. 1-22.

[18] Christopher C. Sentiment Symposium. [Online]. http://sentiment.christopherpotts.net/index.html . [cited May 10, 2015]

[19] Pang B. and Lee L., "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008.

[20] Paul M.J. and Dredze M., "You are what you Tweet: Analyzing Twitter for public health.," *5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, May 2011.

[21] Chew C. and Eysenbach G., "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," , 2010.

[22] Scanfeld D1., Scanfeld V., and Larson EL., "Dissemination of health information through social networks: twitter and antibiotics," in *American journal of infection control.*, vol. 33, 2010, pp. 182-188.

[23] Pak A. and Paroubek P., "Twitter as a Corpus for Sentiment Analysis and Opinion Mining.," in *LREC*, vol. 10, Orsay Cedex, 2010, pp. 1320-1326.

[24] Yang C., Hsin-Yih Lin K., and Chen H., "Emotion classification using web blog corpora. In WI '07," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, Washington, 2007, pp. 275-278.

[25] Read J., "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *In ACL. The Association for Computer Linguistics.*, Stroudsburg, 2005, pp. 43-48.

[26] Go A., Bhayani R., and Huang L., "Twitter sentiment classification using distant supervision," , 2009, pp. 1-6.

[27] Barbosa L. and Feng J., "Robust sentiment detection on twitter from biased and noisy data," in *Proc. 23rd International Conference on Computational Linguistics*, Stroudsburg, 2010, pp. 36-44.

[28] Agarwal A., Xie B., Vovsha I., Rambow O., and Passonneau R., "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Languages in Social Media*, Stroudsburg, 2011, pp. 30-38.

[29] Twitter. Twitter. [Online]. https://dev.twitter.com/streaming/overview . [cited Oct 20, 2015]

[30] Goyal A., Daum III H., and Cormode G., "Sketch algorithms for estimating point queries in NLP," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Stroudsburg, 2012, pp. 1093-1103.

[31] Bifet A. and Frank E., "Sentiment knowledge discovery in twitter streaming data," in *Discovery Science*, Berlin, 2010, pp. 1-15.

[32] Brownlee J. machinelearningmastery. [Online]. http://machinelearningmastery.com/naive-bayes-classifier-scratch-python/ . [cited Oct 20, 2015]

[33] Hunter J. . Matplotlib. [Online]. http://matplotlib.org/ [cited Oct 20, 2015]

[34] Yale Universtiy Department of Statistics. Yale Universtiy Department of Statistics. [Online]. http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm . [cited Oct 20, 2015]

[35] Shaw P. et al., "Intellectual ability and cortical development in children and adolescents," in *Nature*, vol. 440, 2006, pp. 676-679.