A
Dissertation On

Application of Machine Learning in Sentiment Analysis

Submitted in Partial Fulfillment of the Requirement
For the Award of Degree of

Master of Technology

In

Software Technology

By

Sahil Puri
University Roll No. 2K15/SWT/514

Under the Esteemed Guidance of

Dr. Kapil Sharma
Professor, Department of Information Technology, DTU

2015-2019(Jan)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
DELHI – 110042, INDIA

STUDENT UNDERTAKING

Delhi Technological University
(Government of Delhi NCR)
Bawana Road, Delhi- 110042

This is to certify that the thesis entitled **Application of Machine Learning in Sentiment Analysis** done by me for the Major project-II for the achievement of Master of Technology Degree in Software Technology in the Department of Computer Science & Engineering, Delhi Technological University, Delhi is an authentic work carried out by me under the guidance of Professor Dr. Kapil Sharma.

Signature:
Student Name
Sahil Puri
2K15/SWT/514

Above Statement given by Student is Correct.

Project Guide:
Dr. Kapil Sharma , Professor & HOD,
Department of Information Technology, DTU

ACKNOWLEDGEMENT

I would like to express sincere thanks and respect towards my guide **Dr. Kapil Sharma, Professor, Department of Information Technology, Delhi Technological University Delhi.**

I consider myself very fortunate to get the opportunity for work with her and for the guidance I have received from her, while working on this project. Without her support and timely guidance, the completion of the project would have seemed a far. Special thanks for not only providing me necessary project information but also teaching the proper style and techniques of documentation and presentation.

SAHIL PURI
M.Tech (Software Technology)
2K15/SWT/514

ABSTRACT

There is ample amount of statements on social sites which can be inferred with the help of sentiment analysis. It is very beneficial to find the public opinion. Sentiment Analysis involves capturing of user's behavior, likes and dislikes of an individual from the generated web content. There is no concrete definition of "Sentiments", but in general they are considered as thoughts, views and attitude of a person arising mainly based on the emotion instead of a reason. Millions of users use social sites to express their sentiment about brands, services, political and religious views, emotions, beliefs or opinions about things, personalities or places and people they interact with.

This data is mostly unorganized, slangs, etc. and therefore, text analytics and natural language processing are used to extract and classify this data. Any Non-contextual and irrelevant contents are identified and discarded. The classification of sentiments will be performed on this data, which goes as follows: a training data set is created manually and based on this training data set sentiment analysis is performed on the twitter comments. Machine learning such as a hybrid Naive Bayesian classifier is utilised with the lexical dictionary and natural language processing for the sentiment classification

Table of Contents

Chapter One:

INTRODUCTION

1.1 Background

Public opinion is subject to change due to many factors but among those factors Social networking websites now a days have become a prominent tool to shape the public opinion. There are many social net-working websites such as "Twitter" which helps in expressing the view of people from all over the world which is evident from the text that appear on twitter reflect the public views. To understand such views of public the tool of Sentiment Analysis is very helpful. “Sentiments” in the present context can be understood as the individual's thoughts and views which is the resultant of the individual's feelings or emotions rather than the rationality. In recent days it is found that popularity of sentiment analysis has increased as different stakeholders have understood the importance of analyzing the ulterior meaning of statements made on social media.

1.2 Problem Statement

There is ample amount of statements on social sites which can be inferred with the help of sentiment analysis. It is very beneficial to find the public opinion. Sentiment analysis is tested tool to measure the opinions relating to product, performance of movies as shown in figure.1, election etc. It is important to note that public opinion play a significant role in evaluating several things like any place, product, or any person. The opinion can be broadly categorized into either positive, negative or neutral. Thus, with the help of sentiment analysis tool one can easily find the what is general mood with respect to any thing.

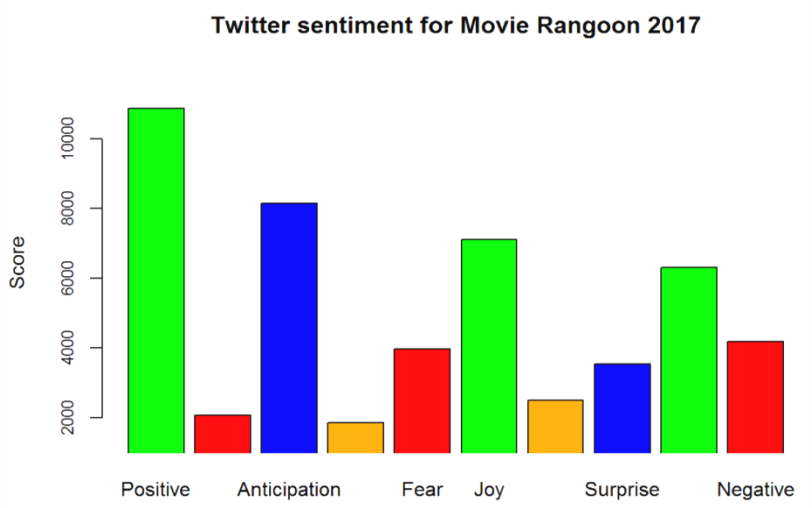

Figure 1: Twitter sentiment analysis of a movie

1.2.1 Analyzing Sentiments

Mining the data related to opinions of people is a challenging task due to the availability of huge data which is getting generated on daily basis as result of communication taking place among the web users. The scope covers a number of discipline as it uses techniques relating to computational linguistics, information retrieval, semantics, natural language processing, artificial intelligence and machine learning. In the present research, the effectiveness of applying machine learning techniques to the sentiment classification problem is examined. Figure 2 shows a typical sentiment analysis process of twitter data. The present problem is a challenge in itself due to its nature of indistinguishable aspect. While classifying the documents in traditional format we simply use the different key words however such application of key words are not possible while classifying the statements related to sentiments because the feelings of people is expressed in indirect way and identifying the keywords that act as a rule is almost impossible. It is therefore we optimise the word/ sentence to learn and thereby near to accurate data

attainment becomes possible on social websites. In such cases, application of data tokenization is done which results in getting the negative or positive values of sentences.

Figure 2: Twitter sentiment analysis process

1.3 Proposed work

There are many steps that are involved while analyse the sentiments. Figure 3 shows the steps involved in generic sentiment analysis approach. The first and foremost step is to gather data from the user tweets. This data is mostly unorganized, slangs, etc. Any Non-contextual and irrelevant contents are identified and discarded.

The classification of sentiments will be performed on this data, which goes as follows: a training data set is created manually and based on this training data set sentiment analysis is performed on the twitter comments. Machine learning such as a hybrid Naive Bayesian classifier is utilised with the lexical dictionary and natural language processing for the sentiment classification. The experiment is setup in python programming language, Twitter APIs and tweepy is used for extraction of tweets. The dataset contained movie tweets. The results are classified as positive, negative and neutral sentiments.

Figure 3: Forward process of sentiment analysis

The sentiment analysis technique generally relies on: a) machine learning based techniques and b) lexicon-based techniques. Python 3.4 is used as programming language for the implementation of experimental setup.

1.4 Motivation

When we as an individual take any decision it is rarely our original thoughts but and it is mostly depend on what we read or hear. Hence one can assume that it is the other person whose way of thoughts are reflected in our own. With widespread use of internet we come across heavy availability of information [1] [2] that has posted in the form of blog or documents but it is rarely easy to find the accurate things as per our requirement. Hence the researchers in the present research has made an attempt the issue in automatic text categorization. Such research will help in organizing the information in such a way can easily be filtered.

It is significant to note that when any person writes any it can be meant for anyone. It is evident from the popularity [3]. Huge texts are being written on Facebook. With an exponential rise in social media usage to share emotions, thoughts and opinion, Twitter has become the gold-mine to analyze brand performance.

These opinions can be clubbed into categories such as negative, positive or neutral. Opinions found on Twitter are casual, honest and informative than what can be picked from formal surveys etc. [3, 4] Millions of users use social sites to express their sentiment about brands, services, political and religious views, emotions, beliefs or opinions about things, personalities or places and people they interact with. As more and more users post their opinions and views, microblogging websites become valuable sources of people's opinions and sentiments. Such data can be efficiently used for various analysis-protocols or monitoring marketing, criminal activities and social studies.

1.5 Problem Statement

The demand of sentiment analysis is raised due to increased requirement of analyzing and structuring hidden information which comes from the social media in the form of unstructured data. Given a set of data containing multiple features and varied opinions, the objective is to extract expressions of opinion describing a target feature and classify it as positive or negative. The greatest challenge of sentiment analysis is to design application-specific algorithms and techniques that can analyze the human language linguistics accurately. We propose a hybrid approach for sentiment analysis that is a combination of a machine learning algorithm (Naive Bayes) and a special lexical dictionary with NLP.

1.6 Thesis Organisation

Chapter 1: In the chapter one, we tried to give general idea of sentiment analysis while describing the methods that has been applied in natural language processing.

Chapter 2: In this chapter, we discussed the previous work done in the field.

Chapter 3: In this chapter we have dealt with the concept of sentiment analysis. It also discusses that why we need this analysis and how do we perform. Levels of sentiment analysis, challenges and the proposed method is also discussed.

Chapter 4: Method of sentiment analysis and Machine learning technique has been elaborated.

Chapter 5: In this chapter Model Social media network: Twitter has been taken into consideration.

Chapter 6: Sentiment analysis classification has been elaborated based on the research.

Chapter 7: This chapter concludes the research.

1.7 Literature Reviews

In this section we summarize the findings from the literature review conducted to understand what metrics and features have worked well and how can they be adopted to sentiment analysis visualization. We also describe other related public tools that perform the similar task [5]. “Sentiment analysis is a very challenging task .Researchers from natural language processing and information retrieval have developed different approaches to solve this problem, achieving promising or satisfying results” [6, 7]. Suchita and Sachin [8], states while comparing the two methods that SVM has shown more in accuracy in comparison with Naive Bayes .

One of the popular work in the field of Twitter sentiment analysis was done by Go et al. in 2009 [9]. They used SVM, MaxEnt and Naive Bayes and reported that SVM and Naive Bayes were equally good and beat MaxEnt. The research basically used distant supervision technique to overcome the problem of manual annotation of large set of tweets. Tweets with ":)" emoticon were considered positive while tweets with ":(" in the message were considered negative. This resulted in two defects. First, emoticons, which were considered important by other works in the area couldn't be used to learn sentiments. Secondly, the authors were unsure if all tweets with ":)" are truly positive or can contain negative or sarcastic sentiments too. Therefore, the dataset used in the experiment was labelled noisy.

The same distant supervision procedure of tagging tweets positive or negative based on the emoticons it mentions was used by Pak and Paroubek [10] also used linear kernel SVM to run the experiment. Although, the results were not reported as accuracy metric. However, highlighting the significance of neutral class, the team also collected neutral tweets. These tweets were strictly objective and were collected from newspapers and magazines.

In this paper [11], the researcher has used the Naive Bayesian classifier in their attempt to analyze the sentence. In their research they tried to show with the help of their experiment by using Naive Bayesian classifier model that large data set can easily be analyzed.

Tumasjan et al. and Bollen et al. [12, 13] employed pre-defined dictionaries for measuring the sentiment level of Tweets.Hu et al. [14] incorporated social signals into their unsupervised sentiment analysis framework. They defined and integrated both emotion indication and correlation into a framework to learn parameters for their sentiment classifier.

In yet another experiment real time data retrieved from two different accounts of politicians were used [15]. In yet another similar research Twitter-streaming application programming interface (API) [16, 17] were used to extract the data. Two sentiment analyzers named SentiWordNet [18, 19] and WordNet [20] were used to find positive and

negative scores . Twitter streaming API was also used to gather data by the authors of [21] for the prediction of the Indonesian presidential elections. The aim was to use Twitter data to understand public opinion .

Fang et al. [22] on the other hand for SVM learning used both general purpose sentiment lexicon along with Domain Specific Sentiment Lexicons. It helped them to get the product aspect.

Zhang et al. And Trinh et al. [23, 24] employed an augmented lexicon-based method which meant for entity level sentiment analysis. It has been stated, "Since the opinions in the twitter are heterogeneous, highly unstructured and along with these it includes positive, negative or neutral in different situation, it is important to analyze the sentiments" [25, 26].

In yet another research, scholar while analyzing the sentiments objective sentences have been ignored [27, 28] .Parikh and Movassate [29] tried the models of Naive Bayes bigram as well as a Maximum Entropy for the classification purpose. Naive Bayes classifiers gave a better result was being concluded in their research.

Chapter Two:

TOOLS FOR SENTIMENT ANALYSIS

2.1 Sentiment Analysis

When the viewpoints of people are being detected automatically with the help of Natural Language Processing (NLP) it is through an analysis which termed as sentiment analysis. What it results into is that it categorise the sentiments in three parts first one to be "positive" second to be "negative" and third to be "neutral" [30]. Sentiment analysis is the task of detecting, extracting and classifying opinions, sentiments and attitudes concerning different topics, as expressed in textual input [31]. SA helps in achieving various goals like observing public mood regarding political movement, market intelligence [31] , the measurement of customer satisfaction [32], movie sales prediction and many more. For this, the opinions are collected from the users, which can be employed for further analysis and improvements. Sentiment analysis provide the comprehension information related to public views, as it analyzes different tweets and reviews. It is a verified tool for the prediction of many significant events such as box office performance of movies and general elections[35].

Sentiment Analysis is the wonder of extricating sentiments or suppositions from surveys communicated by clients over a specific subject, territory or item on the web. The motivation behind sentiment analysis is to naturally decide the expressive heading of client surveys [36].

2.1.1 Why Sentiment Analysis

Most clients utilize social destinations to express their feelings, convictions, perspectives or assessments about different items, administrations, spots or motion pictures. Clients need to see the sentiment of other about an item before getting it. Business associations need to comprehend what clients are stating about their item or administration that an

association is giving, to settle on future choices. With the real advancement in long range informal communication (i.e., Facebook, Twitter, Instagram, LinkedIn, Stumble upon and so forth.,) on the Web, people and enormous affiliations are focusing on popular's assessment for their basic leadership. The assignment of mining feeling data on sites isn't simple one, as a result of tremendous number of sites at present and as yet populating and due to absence of institutionalized philosophy to do likewise. In addition, the content corpora present on sites establish both futile and valuable information that is required for the analysis. There are different factors also like human mental limit and physical impediment that make people unfit to examine huge measure of information. Consequently, a mechanized feeling mining is required which will in the long run assistance people in sentiment analysis.

A sentiment grouping should be possible at Document level, Sentence level and Aspect or Feature level [37]. In Document level the entire archive is utilized as an essential data unit to arrange it either into positive or negative class. Sentence level sentiment arrangement orders each sentence first as abstract or goal and after that characterizes into positive, negative or impartial class. Angle or Feature level sentiment arrangement manages recognizing and extricating item includes from the source information [38-39].

For the most part, there are different methodologies in sentimental analysis, for example, thinking about representative technique, utilizing lexicon word reference or by AI strategy. In emblematic learning system, which is ordered by some taking in techniques, for

example, gaining from relationship, revelation, precedents and from root learning. Figure 1 speaks to the fundamental ideas associated with sentiment analysis. In AI strategy it utilizes unsupervised adapting, pitifully administered learning and managed learning. Alongside lexicon based and etymological technique, AI will be considered as one of the primarily utilized methodology in sentiment arrangement. Figure 2 demonstrates a run of the mill stream chart of sentiment analysis order.

2.1.2 SA: How It Works?

A. **Subjectivity/Objectivity-** To perform sentiment analysis we first need to identify the subjective and objective text. Only subjective text holds the sentiments. Objective text contains only factual-information.

   Examples:

   1. ***Subjective:*** Inception is a superb movie. (this sentence has a sentimental opinion "superb", which talks about the movie and the writer's feelings, thus it is subjective).

   2. ***Objective:*** James Cameron is the director of titanic. (this sentence has no sentiment, it is a fact, a general information rather than an opinion or a view of some individual thus it is objective).

B. **Polarity-** Further subjective text can be classified into 3 categories based on the sentiments conveyed in the text.

   Examples:

   1. *Positive:* I love the new Transformers movie.

   2. *Negative:* The actions and graphics of the movie were awful.

3. *Neutral:* I usually get tired by the evening. (this sentence has user's views, feelings, hence, it is subjective but as it does not have any positive or negative polarity so it is neutral.)

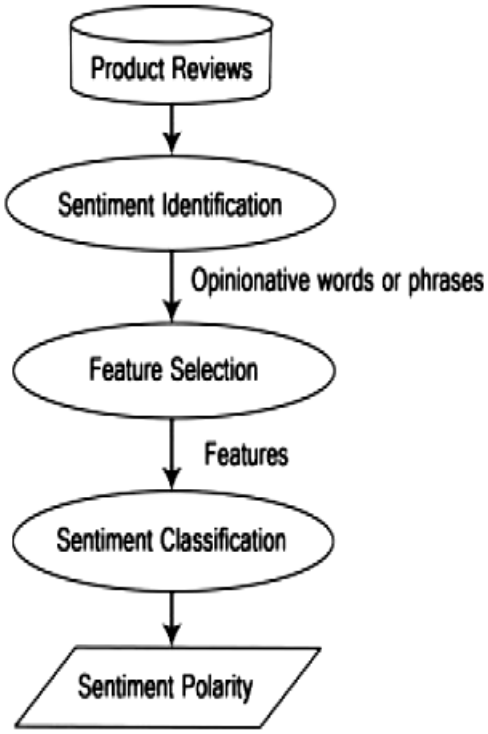

Figure 4: Flow diagram of classification of sentiment analysis

Though certain words infer similar meaning ad usually are used interchangeably but has different meaning.

Opinion: A conclusive remark that ignites other to speak out their mind

View: subjective Opinion that has different meaning

Belief: deliberate acceptance and intellectual assent

Sentiment: opinion representing one's feelings

2.2 Natural language processing

Natural language processing deals with advancement of frameworks which assist in peoples' communication in the language that they generally use [40]. With regard to Association for Computational Linguistics, principally it manages the scope of procedures. Here when we talk of figuring, examining and communicating to naturally happening writings at staggered investigation of languages. This is done because the process of machine functions like that of human language when it comes to applications. NLP calculations depend profoundly on machine learning with the dominant part being factual. More established execution of language-preparing undertakings regularly required hard coding of enormous arrangement of standards [41].

By utilizing machine learning, we can utilize ordinary learning calculations normally in factual derivation, to learn governs by investigating substantial corpora of genuine models.

2.2.1 Kinds of Natural Language Processing are:

1. *Morphological preparing:*

In this condition of language preparing, strings are broken into sets of tokens relating to the words, sub-words and accentuation frames. Typically, a modification happens by including prefixes or postfixes as well as might be by different changes.

2. *Syntax and semantic investigation:*

A processor that does diff erent works principally dependent on syntax and semantic investigation. There are two employments of syntax investigation. One is to check if a sentence is all around framed and the other is to break it into a structure that gives syntactic connection between them. The equivalent can be accomplished by a parser utilizing a lexicon of word definitions and an arrangement of syntax rules. A straightforward dictionary has the syntactic class of each word, the principles are depicted by the sentence structure which mean how they can join expressions of different kinds.

3. *Pragmatic investigation:*

Translating the consequences of semantic investigation considering from the purpose of perspective of a specific circumstance is called pragmatic investigation.

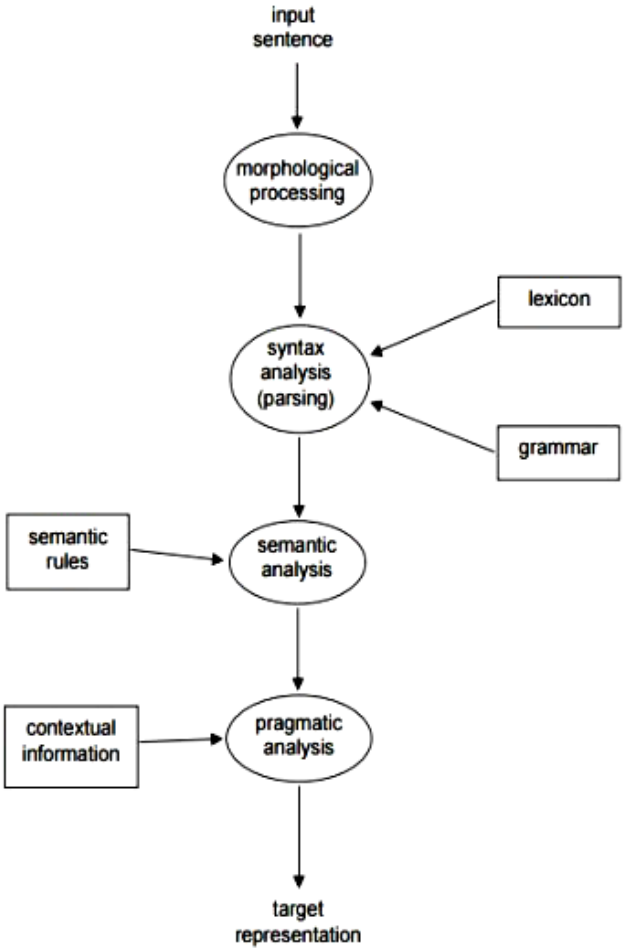

Figure 5: Flow-diagram of NLP steps

2.3 Sentiment analysis levels

Sentiment analysis generally categorised as mentioned below [42].

a) Document Level:

It oversees marking particular chronicles with their inclination. In Document level the whole file is orchestrate either into positive or negative class. Here, the essential

objective getting satisfied is that of in portraying complete supposition document imparting either a positive or negative assessment. This dimension of examination acknowledge that each report conveys determinations on a singular component.

b) Phrase or Sentence Level:

Expression level Sentiment Analysis oversees marking solitary sentences with their specific decision polarities. Sentence level supposition gathering orders sentence into positive, negative or unprejudiced class. Here, the key endeavor is to check in every case about sentences being imparting a meaning which is positive, negative, or unprejudiced sentiment. The present dimension in case of examining depend upon subjectivity portrayal. The perceived targeted sentence expressed honest clue from conceptual sentences. Record level and the sentence level examinations don't discover what correctly people delighted in and despised.

c) Aspect Level:

The Aspect level arrangements with marking each word with their sentiment and, additionally distinguishing the substance towards which the sentiment is coordinated. Perspective or Feature level sentiment order worries with distinguishing and removing item includes from the source information. Methods like reliance parser and talk structures are utilized in this. This strategy level performs better grained analysis. Rather than seeing language builds (archives, passages, sentences, conditions or expressions), viewpoint level straightforwardly takes a gander at the feeling itself.

d) Word Level:

Latest works have utilized the earlier extremity of words and expressions for sentiment characterization at sentence and record levels Word sentiment order use for the

most part descriptors as highlights yet verb modifiers. The two strategies for consequently commenting on sentiment at the word level are:

(1) Dictionary-Based Approaches

(2) Corpus-Based Approaches

Figure 5 demonstrates the dimensions of sentiment analysis. Figure 6 and 7 show graphical examination of a couple of techniques for twitter characterization.

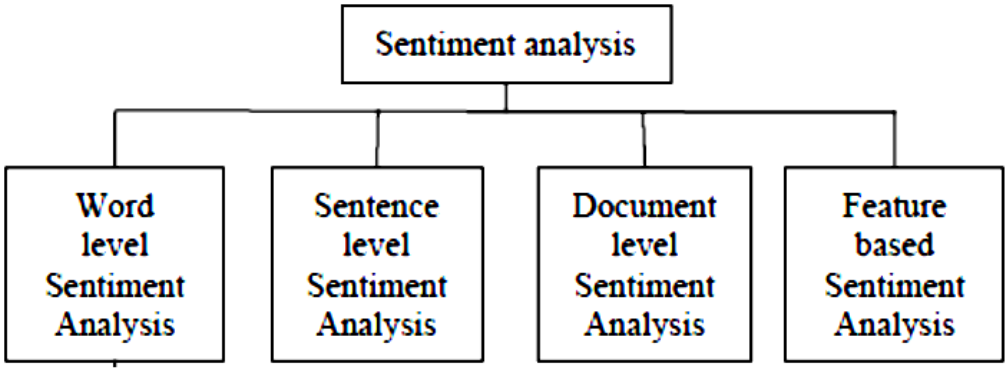

Figure 6: Sentiment analysis levels

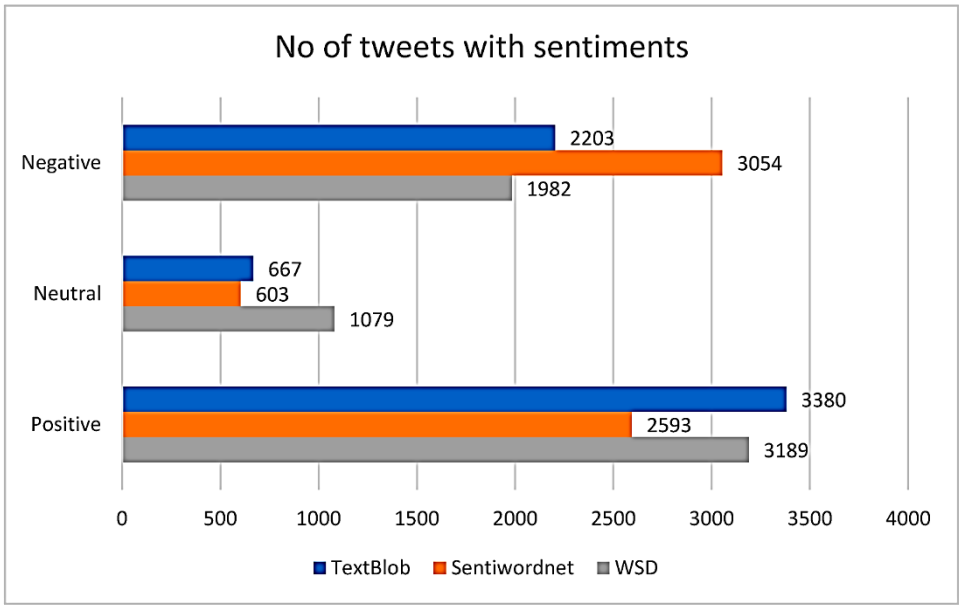

Figure 7: Sentiment classification based on word assessment

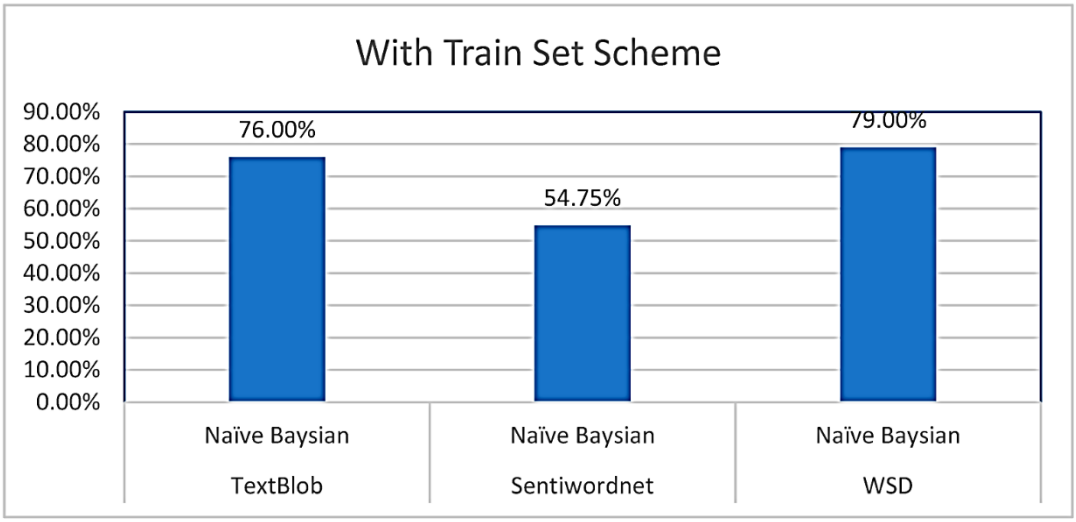

Figure 8: Performance of Naive Bayesian classifier with different analyzers

2.4 Challenges in Sentiment Analysis

Sentiment Analysis is a very challenging task. Following are some of the challenges faced in Sentiment Analysis of Twitter.

1. Segmenting the subjective aspect in text: Emotional parts speak to feeling bearing substance. A similar word can be treated as emotional in one case, or objective in some other. This makes it hard to recognize the subjective bits of content.

2. Domain dependence [24]: Multiple meaning of a word is yet another issue. More often the words use to have its meaning when it is used in a certain sentence. Therefore, it is the domain that decides the meaning.

3. Sarcasm Detection: The problem with a Sarcastic sentence is that positive words are used but ultimately bears a negative meaning.

2.5 Proposed Method

The proposed model (figure 9) is a prediction system that predicts the behavior of user, specifically twitter account holders on twitter, through which the behavior of each

user and groups are predicted or their character is determined using their tweets. This work implements python and machine learning with Naive Bayes classifier for the purpose.

The system works as follows: the input of the system contains twitter commends that are collected from twitter tweets and after the data collection these tweets are pre-processed. The data will be well cleaned and irrelevant data might be removed after pre-processing the cleaned data then subjected to sentiment analysis in order to calculate the sentiment of the tweets. The sentiment is calculated by using a training dataset which was deliberately created, based on the training dataset polarity is being calculated for each tweet. Finally, based on the polarity of these tweets they are classifies into positive, negative or neutral.

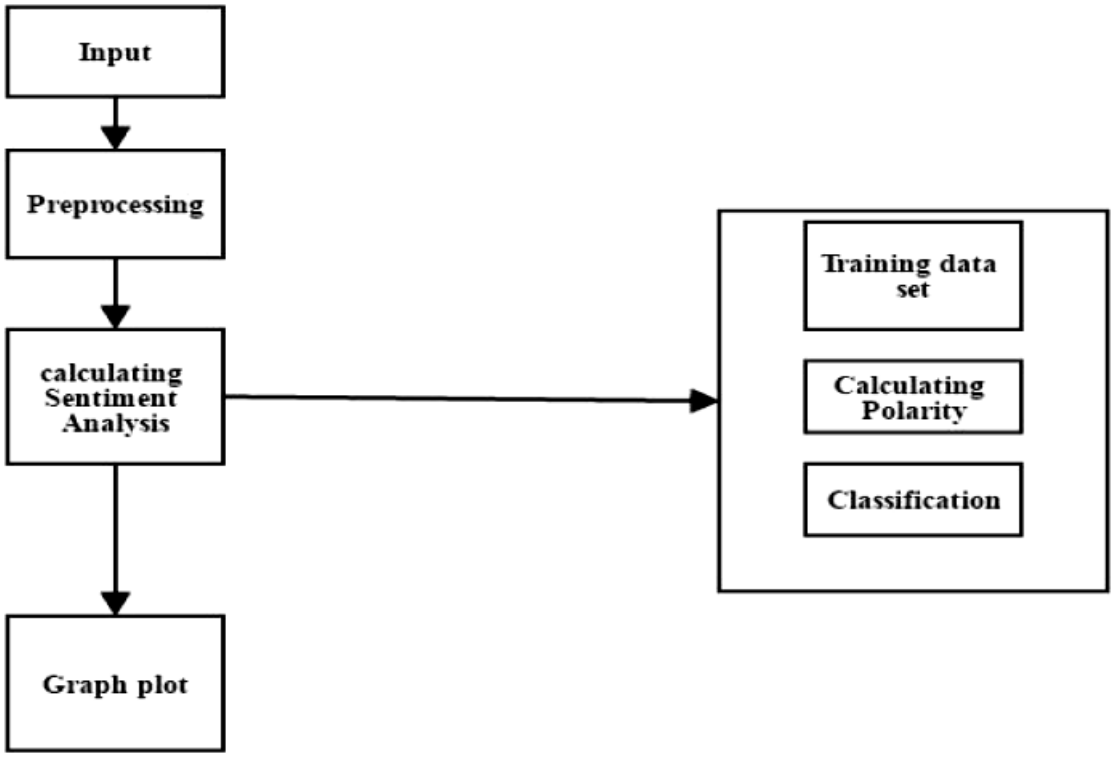

Figure 9: Architecture of sentiment analysis

Chapter Three:

METHODS FOR ANALYSIS

3.1 Methods for Sentiment Analysis

There are two main techniques for sentiment analysis: machine learning based and lexicon based. Machine learning usually takes the supervised or unsupervised approaches (as shown if figure 10). Few research studies have also combined these two methods and gain relatively better performance.

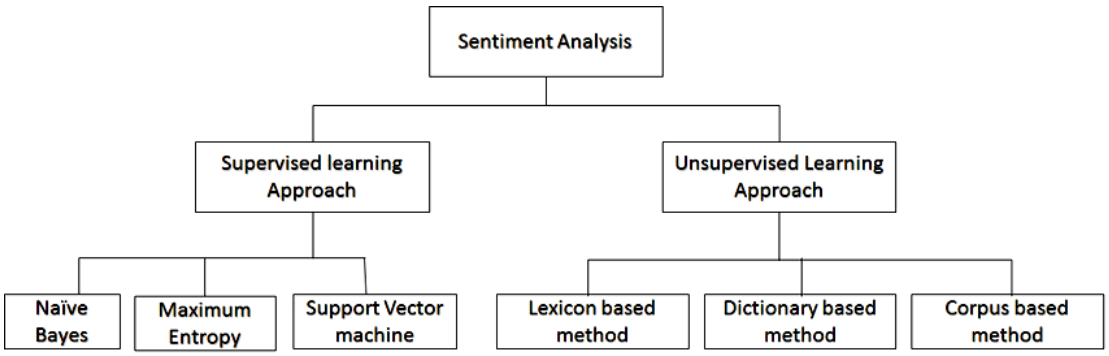

Figure 10: Methods of sentiment analysis

Table 1: Types of sentiment classifiers (with advantages and disadvantages)

Sentiment Classification Approaches		Features / Tecniques	Advantages and Limitations
Machine learning	Bayesian Networks Naive Bayes Classification Maximum Entropy Neural Networks Support Vector Machine	Term presence and frequency Negations Part of speech information Opinion words and phrases	Advantages the ability to adapt and create trained models for specific purposes and contexts Limitations the low applicability to new data because it is necessary the availability of labeled data that could be costly or even prohibitive
Lexicon based	Dictionary based Novel Machine Learning	Manual construction, Corpus-based	Advantages wider term coverage Limitations finite number of words in the

	Corpus based Ensemble Approaches	Dictionary based	lexicons and the assignation of a fixed sentiment orientation and score to words
Hybrid	Machine Learning and Lexicon based	Sentiment lexicon constructed using public resources for initial sentiment Detection. Sentiment words as features in machine learning method.	Advantages lexicon/learning symbiosis, the detection and measurement of sentiment at the concept level and the lesser sensitivity to changes in topic domain Limitations noisy reviews

3.1.1 Machine Learning Technique

AI is a part of man-made reasoning that enables frameworks to take in consequently and improve themselves from the experience without being expressly modified or without the mediation of human. Its principle point is to cause PCs to gain consequently from the experience. Figure 11 indicates various degrees and execution of AI.

The AI approach appropriate to sentiment analysis for the most part has a place with managed grouping. A regulated learning classifier utilizes the preparation set to learn and prepare itself concerning the separating properties of content, and the exhibition of the classifier is tried utilizing test dataset. In this, two arrangements of archives are required: preparing and a test set. A preparation set is utilized by a programmed classifier to gain

proficiency with the separating qualities of reports, and a test set is utilized to check how well the classifier performs.

Different AI systems have been received to characterize the surveys. AI begins with gathering preparing dataset. A few AI calculations like Maximum Entropy (ME), Naive Baye's (NB) and Support Vector Machines (SVM) are generally utilized for characterization of content (tweets). These calculations have created victories. Table 1 speaks to a correlation table with focal points and impediments of different machine classifiers.

3.1.2 Deep Learning

Deep Learning was firstly proposed by G.E. Hinton in 2006 and is the part of machine learning process which refers to Deep Neural Network. Neural network is influenced by human brain and it contains several neurons that make an impressive network (figure 12). Deep learning is very effective in learning robust features in a supervised or unsupervised fashion. Deep learning networks are capable for providing training to both supervised and unsupervised categories [43].

Deep learning has been recently extended to include different other networks such as CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks), Recursive Neural Networks, DNN (Deep Neural Networks), DBN (Deep Belief Networks) and many more. Neural networks are very beneficial in text generation, vector representation, word representation estimation, sentence classification, sentence modeling and feature presentation.

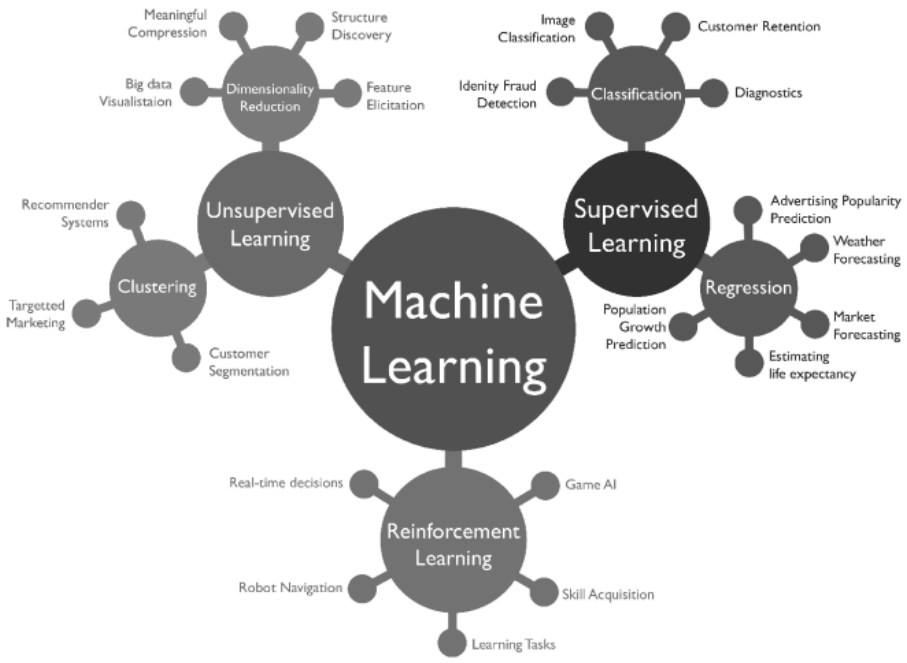

Figure 11: Branches and scopes of Machine learning

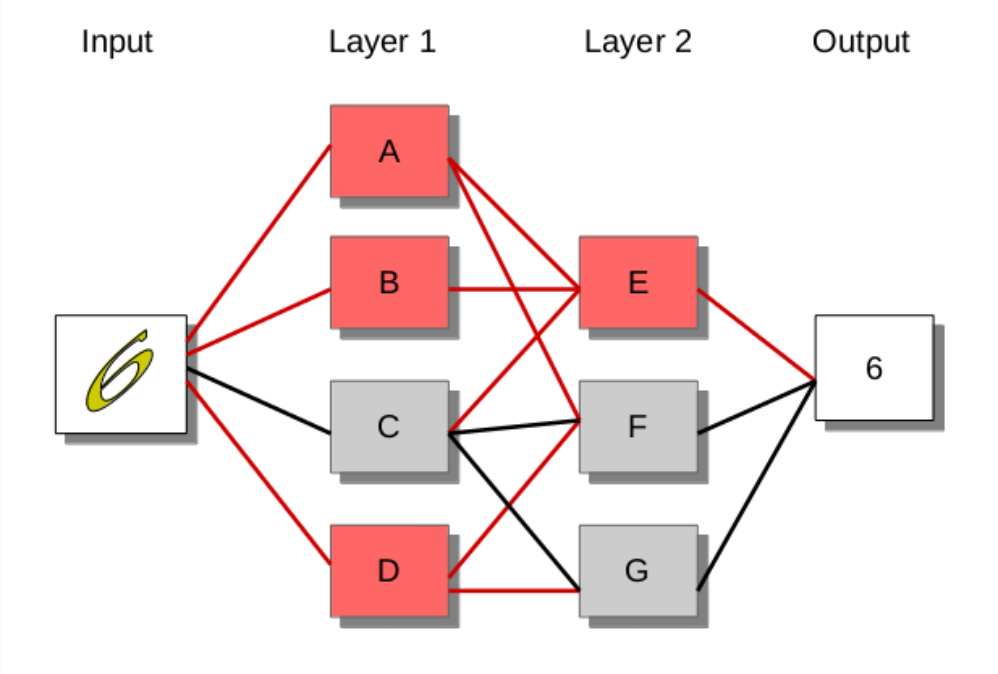

Figure 12: A 2 layered Neural Network

3.1.3 Supervised Learning

Machine learning starts with collecting training dataset. The next step is to train a classifier on the training data. Once a supervised classification technique is selected, an important decision to make is feature selection. In case of supervised machine learning technique we find that there is an association with marked feature uses. In this case, classification function hold some learning function from the experiment along with its input and output. Preparing informational index incorporates set of preparing precedents; every model comprises of couple of an information just as anticipated yield. The accuracy of predictions by the computer during training is also analyzed.

3.1.4 Unsupervised Learning

For this situation, no such preparational input is given, leaving PCs to discover the yield without any input. Unsupervised learning is for the most part connected on value-based information. It is utilized in more intricate undertakings. It utilizes another methodology of cycle known a profound figuring out how to land at a few ends. A few models for unsupervised learning approach are bunch examination, desire amplification calculations. These calculations utilize Dictionary based way to deal with assemble nostalgic content.

3.1.5 Naive Bayes Method

It is a probabilistic classifier and is essentially utilized when the extent of the preparation set is less. In machine learning it is in group of test probabilistic classifier dependent on Bayes hypothesis. The restrictive likelihood that an occasion X happens given the proof Y is controlled by Bayes rule by:

$$P(X/Y) = P(X)\, P(Y/X)/P(Y)$$

Thus, to discover the supposition the condition is changed into the beneath:

P(Sentiment/Sentence) = P(Sentiment)P(Sentence/Sentiment)/P(Sentence)

P (sentence/slant) is computed as the result of P (token/opinion), which is detailed by:

Tally (Thistokeninclass) + 1/Count (Alltokensinclass) + Count (Alltokens)

Here 1 and tally of all tokens is called include one or Laplace smoothing.

3.1.6 k-Nearest Neighbor method and weighted k-Nearest Neighbor method

K-NN technique depends on the way that the arrangement of an occurrence will be to some degree like those adjacent it in the vector space. Facilitate some gathering, inspected on weighted k-Nearest Neighbor strategy, in which they gave weightage to those components in the preparation set and they utilized these weights for their count of assumption of content in word by word way [44].

Inspiration Score = (1∑j score (pos) + 1∑k score (neg))/1∑s most extreme score

Here *s=j+k,* i.e. check of both positive and negative together. In weighted k-NN technique, they as a matter of first importance, tokenize the sentences and expelled the prevent words from the tweets they have brought. A positive score is doled out to every survey after the main parse. This is passed for second parsing and a contribution of nonpartisan survey is given. Utilizing this the score is changed whenever required. It is improved the situation better positivist assurance and a yield document comprising of survey ID and its positive score is resolved.

3.1.7 Lexicon based techniques

These procedures depend on decision trees, for example, k-Nearest Neighbors (k-NN), Conditional Random Field (CRF), Hidden Markov Model (HMM), Naive Bayesian Classifier (NBC), Single Dimensional Classification (SDC) and Sequential Minimal Optimization (SMO) that are identified with techniques of conclusion classification. Vocabulary Based methods chip away at a suspicion that the aggregate extremity of a sentence or reports is the total of polarities of the individual expressions or words. In the

2011-12 ROMIP and RCDL course the dictionary-based strategy was utilized [45, 46]. This strategy depends on passionate research for slant investigation lexicons for every space. Next, every area lexicon was renewed with evaluation expressions of suitable preparing gathering that have the most elevated weight, ascertained by the strategy for RF (Relevance Frequency) [47].

In unsupervised strategy, order is finished by looking at the highlights of a given content against assessment vocabularies whose opinion esteems are resolved preceding their utilization. Estimation dictionary contains arrangements of words and articulations used to express individuals' emotional sentiments and conclusions. For instance, begin with positive and negative word dictionaries, examine the archive for which assumption need to discover. At that point if the archive has more positive word dictionaries, it is certain, else it is negative. The dictionary-based procedures to Sentiment investigation is unsupervised learning since it doesn't require earlier preparing with the end goal to characterize the information. The essential strides of the vocabulary-based systems are laid out beneath:

1. Pre-process each text (i.e. remove HTML tags, noisy characters).

2. "Initialize the total text sentiment score: $s \leftarrow 0$."

3. Tokenize text. For each token, check if it is present in a sentiment dictionary.

    a) If token is present in dictionary,

        i. If token is positive, then $s \leftarrow s + w$.

        ii. If token is negative, then $s \leftarrow s - w$.

4. Look at total text sentiment score s,

    a) If s > threshold, then classify the text as positive.

    b) If s < threshold, then classify the text as negative.

There are three methods to construct a sentiment lexicon: manual construction, corpus-based methods and dictionary-based methods. The manual construction of sentiment lexicon is a difficult and time-consuming task. In dictionary-based techniques the idea is to first collect a small set of opinion words manually with known orientations,

and then to grow this set by searching in the WordNet dictionary for their synonyms and antonyms.

3.1.8 N-gram Sentiment Analysis

In the fields of phonetics and probability, a n-gram is a coterminous succession of n things from a given grouping of content or discourse. The things can be phonemes, syllables, letters, words or base sets as per the application. The n-grams normally are gathered from a content or discourse corpus. At the point when the things are words, n-grams may likewise be called shingles. In this they are thinking about the sentence all in all [48]. They are making utilization of four kinds of vocabularies to be specific conclusion express vocabulary, estimation quality dictionary, vocabulary with viewpoints and exemption lexicons.

3.1.9 Multilingual Sentiment Analysis

Nowadays, with so many choices available for the users there is a possibility that users may bring in many languages into use while sharing the views. In such scenario it becomes obvious for any researcher to bring the consideration of different languages. Some researches [49, 50] explained methods within multilingual framework to carry out the task of determining the polarity of the text. It is done using several Natural Language Tool Kits (NLTK). In this, language is identified first using language models. After identification, the language is translated to English using standard translation software.

3.1.10 Maximum Entropy Classifier

In case of Maximum Entropy (ME) classifier, a set of weights are parameterized. This further can further be utilized for mixing the features which can be produced with the help of features through encoding. The encoding maps each match of list of capabilities and name to a vector. ME classifiers have a place with the arrangement of classifiers known as the exponential or log-straight classifiers, since they work by extricating some arrangement of highlights from the information, consolidating them directly and after that utilizing this total as example. On the off chance that this strategy is done in an unsupervised way, at that Point insightful Mutual Information (PMI) is made use with the

end goal to discover the co-event of a word with positive and negative words. The ME Classifier is one of the models which don't expect the autonomous highlights [51].

$$P_{ME}(c \mid d, \lambda) = \frac{\exp[\sum_i \lambda_i f_i(c, d)]}{\sum_c \exp[\sum_i \lambda_i f_i(c, d)]}$$

Where c is the class, d is the tweet and λ_i is the weight vector. The weight vectors decide the importance of a feature in classification.

Chapter Four:

PROPOSED APPROACH

4.1 The Model on social media network: *Twitter*

4.1.1 Twitter Sentiment Analysis

Twitter is an online person to person communication administration and microblogging administration that empowers its clients to send and peruse content based messages called "tweets". Tweets are freely obvious of course, however senders can confine the message conveyance to a restricted group. Twitter is one of the biggest microblogging administrations having more than 500 million enlisted clients. Measurements uncovered by the Info-designs Labs propose that on regular schedule around a large portion of a billion tweets are conveyed. There is an enormous mass of individuals utilizing twitter to express slants, which makes it a fascinating and testing decision for supposition examination. At the point when so much consideration is being paid to twitter, why not screen and develop strategies to examine these notions [52]. Ongoing examinations have demonstrated [53,54] that with Twitter it is conceivable to get individuals' understanding from their profiles as opposed to customary methods for acquiring data about discernments.

We zeroed on the data available on Twitter due to more authentic accounts of the people. It is assumed that the views expressed on twitter are far more honest and informative than other modes of surveys (figure 13). Also, in relation to Facebook, Twitter confines clients to give their minimized and finish assessments in 280 characters only [55].

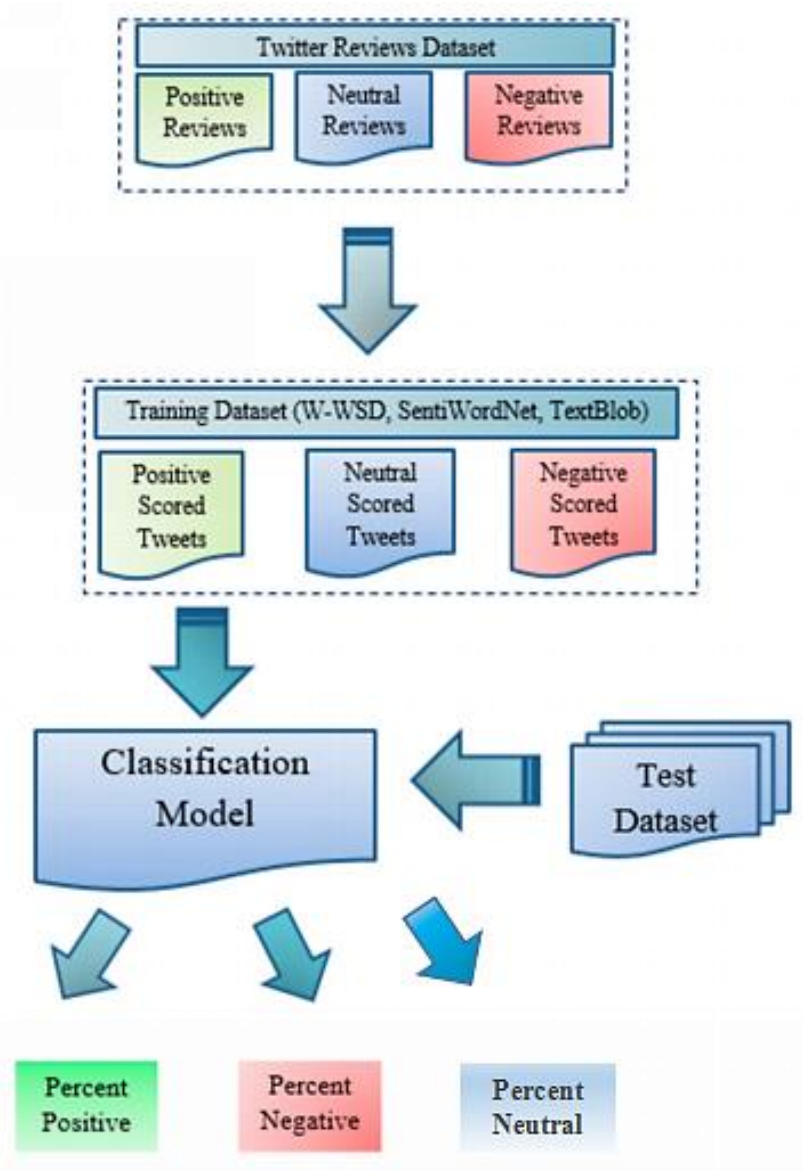

Figure 13: Flow diagram of sentiment analysis framework

4.2 Objective:

It is worth noting that the task of understanding the sentiment in a tweet is more complex that of any well formatted document. Tweets do not follow any formal language structure, nor do they contain words from formal language (i.e. no-vocabulary words). Often, punctuations and symbols are used to express emotions (smileys, emoticons etc.).

Classify tweets in either as positive sentiment or negative sentiment using hybrid techniques and NLP under python program and check the performance.

In this work, we propose a Hybrid approach which is the combination of a machine learning algorithm (Naive Bayes), natural language processing techniques and a special lexical dictionary (SentiWordNet / *www.sentiwordnet.com*) to understand the patterns and characteristics of tweets and predict the sentiment (if any) they carry. Specifically, we build a computational model that can classify a given tweet as either positive, negative or neutral based on the sentiments it reflects. A positive and negative class would contain polar tweets expressing a sentiment. We crept multi-sized datasets comprising of around 4 million tweets with an assortment of most well known watchwords and so on., for the preparation and testing purposes. We test the proposed Hybrid Naive Bayes approach utilizing Natural language Toolkit and see that it outflanks the current methodologies conveying aggressive outcomes having 98.59% precision.

The purpose of finalising the Twitter are as follows.

- Twitter is an Open access social network .
- Twitter is an Ocean of sentiments
- Twitter provides user friendly API making it easier to mine sentiments in real-time.

4.3 Data collection:

Twitter allows researchers to collect tweets by using a Twitter API. One must have a twitter account to obtain twitter credentials (i.e. API key, API secret, Access token and Access token secret) which can be obtained from twitter developer site. Then install a twitter library to connect to the Twitter API. Twitter has developed its own language conventions. The following are examples of Twitter conventions.

a) “RT” is an acronym for retweet, which indicates that the user is repeating or reposting.

b) “#” stands for hashtag is used to filter tweets according to topics or categories.

c) “@user1” represents that a message is a reply to a user whose user name is “user1".

d) Emoticons and colloquial expressions or slang languages are frequently used in tweets.

4.4 Proposed Machine Learning Model:

4.4.1 Hybrid Naive Bayes

It is widely known across world that the method of lexical along with machine learning are best in terms of speed and accuracy respectively. The speed in case of lexical approach happens due to features which are predefined. Having a dictionary to refer at runtime reduces the time consumption almost exponentially. To improve performance of lexical approaches the feature set must be increased drastically, i.e. a very large dictionary of variety of words with their frequencies must be provided at runtime. This increases the overhead of the system and hence the performance suff ers. Thus, there is a constant trade-off between Performance vs Time. On the other hand, machine learning approaches employ recursively learning and tuning of their features, given large input datasets, improves its performance way beyond any lexical approach can achieve. However, due to this runtime performance tuning and learning the system undergoes drastic fall in time constraints.

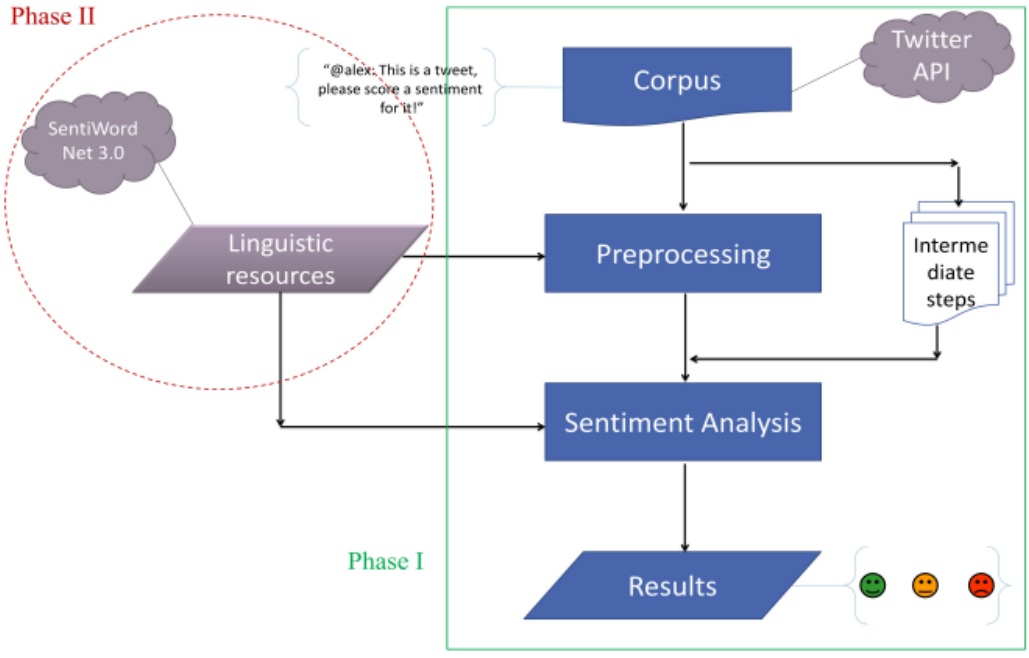

Figure 14: System architecture of hybrid naive bayes approach.

A Naive Bayesian classifier is one of the familiar supervised learning techniques which are frequently used for classification purpose. Figure 14 represents a typical architecture of hybrid Naive Bayes approach. Their classifier is named as naive since it considers the contingency that are actually-linked are not depending on the further. Let d be the tweet and c* be a class that is assigned to d, where

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)},$$

where *P(d)* plays no role in selecting c^*. To estimate the term P(d | c), Naive Bayes decomposes it by assuming the features f_i are conditionally independent given *d*'s class:

$$P_{NB}(c \mid d) = \frac{(P(c))\sum_{i=1}^{m} p(f \mid c)^{n_i(d)}}{P(d)}$$

These stated equations, where in *"f"* stands for "feature" and feature count (f_i) is denoted with $n_i(d)$ and is present in d which represents a tweet. The meaning of m here is the number of feature. Parameters *P(c)* and *P(f|c)* are computed through maximum likelihood estimates, and smoothing is utilized for unseen features. To train and classify using Naive Bayes Machine Learning technique, we can use the Python NLTK library. Despite its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, Naive Bayes-based text categorization still tends to perform surprisingly well.

4.4.2 Support Vector Machines

In the event of conventional content categorization, Support vector machines have been appeared to be exceedingly eff ective. When all is said in done it beats other machine learning strategies. They are extensive edge, instead of probabilistic, classifiers, as opposed to Naive Bayes and MaxEnt. In the two-classification case, the essential thought behind the preparation system is to find a hyperplane (figure 15), spoken to by vector, that not just isolates the archive vectors in a single class from those in the other, however for which the detachment, or edge, is as vast as could reasonably be expected.

$$\vec{w} := \sum_j \alpha_j c_j \vec{d}_j, \quad \alpha_j \geq 0,$$

Where the α j 's are acquired by tackling a double improvement issue. Those dj with the end goal that α j is more noteworthy than zero are called support vectors, since they are the main archive vectors adding to. Classification of test examples comprises essentially of figuring out which side of's hyperplane they fall on.

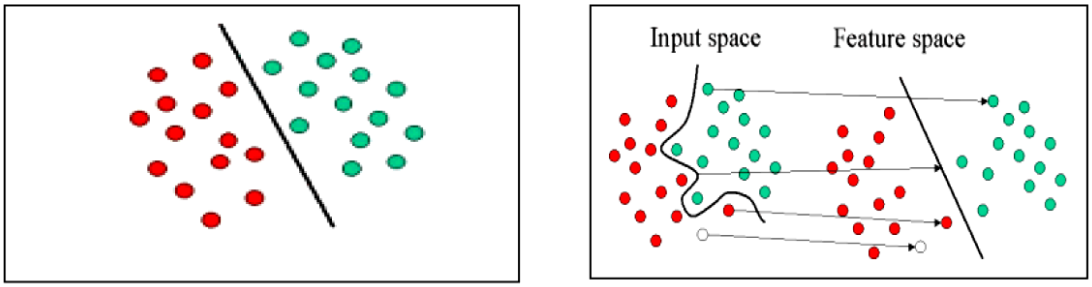

Figure 15: SVM in linear classification

Chapter Five:

CLASSIFICATION STEPS AND RESULTS

5.1 Sentiment Analysis classification

5.1.1 Document-level of sentiment analysis:

In layman understanding the opinions can be understood as the feeling or perception of an individual regarding anything. There are multiple ways to express their sentiments. It can be straightforward Yes or No. It can also be subjectively examined to find the actual meaning of sentiments. To measure the actual viewpoints, review framework can be utilized wherein on one side the rating of 4 or 5 can be assumed as yes and 1 or 2 can be assumed as No.

5.1.2 Sentence-level of sentiment analysis:

This strategy is used to give valuable information when we seek on the grounds that the extremity of sentence will be made flawless. In this level of conclusion investigation experience those sentences which contain assessments and gives audits as if it is negative or positive.

5.2 Preprocessing

When we collect the statements from twitter it use to be the combination of "urls", along with various data of non-sentimental type such as hashtags "#", annotation "@" and retweets "RT". To obtain n-gram features, we must first tokenize the text input. Tweets pose a problem for standard tokenizers designed for formal and regular text. The following figure 16 displays the various intermediate processing feature steps. Figure 17 shows block diagram of the classification based on dictionary.

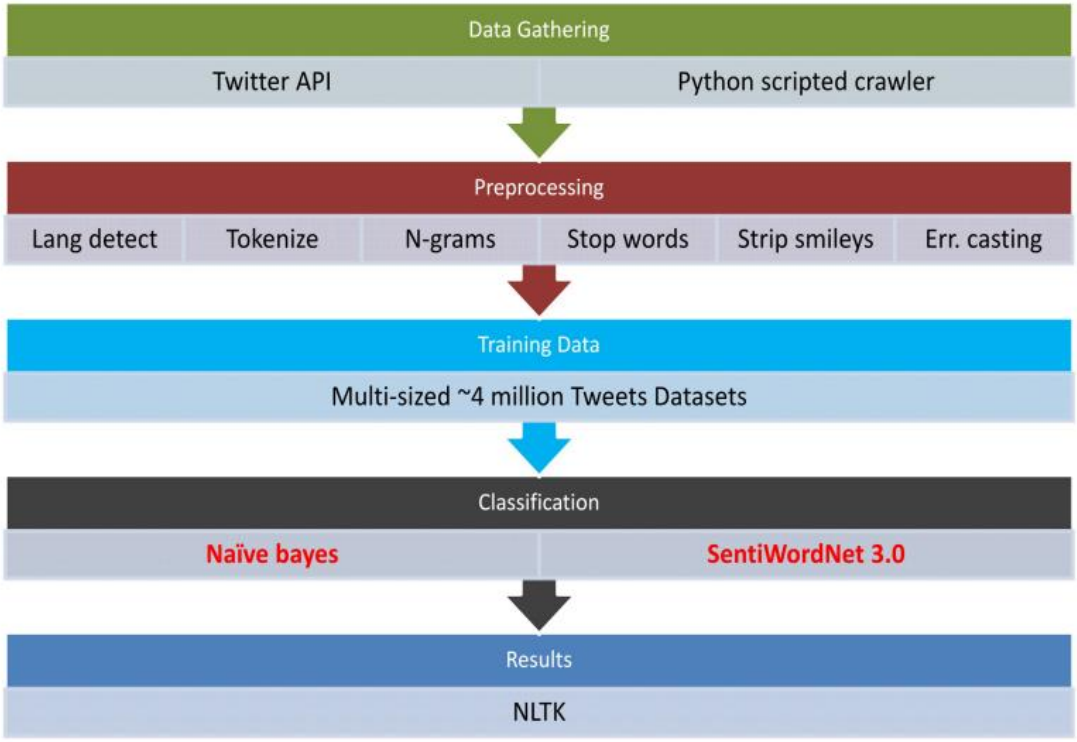

Figure 16: Process steps followed by hybrid Naive Bayes.

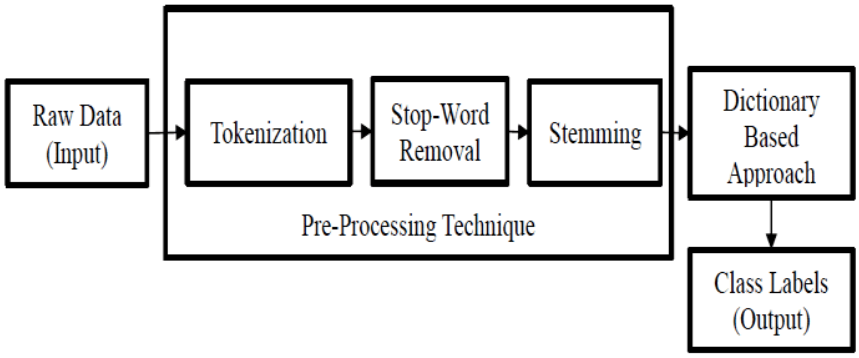

Figure 17: Classification process using Dictionary Based Approach

5.3 Data collection:

There is a requirement of Twitter data for the purpose of classifying the classifier in case of Twitter API. For this purpose, we make use of API's twitter provides. Twitter

provides two API's; Stream API and REST API. The difference between Streaming API and *REST APIs are:*

1) Streaming API supports long-lived connection and provides data in almost real-time. The REST APIs support short-lived connections and are rate-limited (one can download a certain amount of data [*150 tweets per hour] but not more per day).

2) REST APIs allow access to Twitter data such as status updates and user info regardless of time. However, Twitter does not make data older than a week or so available. Thus, REST access is limited to data Twittered not before more than a week. Therefore, while REST API allows access to these accumulated data, Streaming API enables access to data as it is being twittered.

5.4 Polarity Calculation and Sentiment Analysis

There is a very large set of data available resembling the emotions of people on internet that too in a much unstructured format. In such scenario, analysis provided by Sentiment analysis will be very helpful. In this analysis, three classes are utilized one is negative, second is positive and the last is neutral. All the viewpoints are judged by giving the score in the range of −1 to 1. Consistently, all the negative words are given -1 score whereas all the neutral words are given 0 and all the positive are given + 1. A score of subjectivity appointed to every statement depends on whether it is speaking to a subjective importance or objective meaning; the scope of subjectivity score is additionally from 0 to 1 where an incentive close to 0 speaks to objective and close to 1 subjective. For distinguishing the extremity and subjectivity of political audits, and to give an unmistakable perspective of the most precise analyzer for the extremity and subjectivity adding machine, we utilized Textblob and SentiWordNet analyzers.

Table 2 demonstrates the test aftereffects of three feeling analyzers in which we can see the slant relegated by every analyzer. Capacity 3 recorded beneath figures the subjectivity and extremity of handled tweets utilizing every assessment analyzer (SentiWordNet, W-WSD, TextBlob) with the utilization of Python code.

Table 2: Comparison of sentiment scores of different analyzers.

Sentiment Analyzer	Tweet	Sentiment Score
W-WSD	'Right move at wrong time #JIT'	Negative
TextBlob	'Right move at wrong time #JIT'	Negative
SentiWordNet	'Right move at wrong time #JIT'	Positive

In case when one think of developing model, one can apply the algorithm of machine-learning which is supervised, Naive Bayes particularly on the dataset of training. Steps of the analyzers validation through Naive Bayes can be viewed in the below section. These machine-learning algorithms (hybrid Naïve Bayes) were applied on the training set to build an analysis model. Based on the model constructed for each analyzer, the test set was evaluated. After test set evaluation, we recorded the accuracy of another analyzer under each model.

5.5 Naive Bayes Classifier Execution

The first step is to create the data files of the classifier wherein the below mentioned process need to be done.

1) Tweet file will be created which will have the sentiment analyzer.

2) For every analyzer, a model is created with the help of training set file.

Once first step is complete, the second step is to execute the model on test set wherein below mentioned process need to be done.

1) We first load the test set file.

2) Apply the String to Word Vector filter with following parameters: IDF Transform: true, TF Transform: true, stemmer: Snowball Stemmer, stop words Handler: rainbow, tokenizer: Word Tokenizer.

3) Finally, we execute the model on the test set.

4) In the end we save results in the output file.

Once we come up with new words, they are attached in seed-list. There upon we go for another iteration. It ends only we are not left with any word. [8]. opinion words share indistinguishable introduction from their equivalent words and inverse introductions as their antonyms.

5.6 Experimental Setup

5.6.1 Recommended

Operating system Windows XP/7/8/10,Linux (Ubuntu 12.04 or above)

Processor type C2D/i3/i5/i7 (32/64 bits)

Min. Memory (RAM) $\geq$ 4 GB

Min. HDD space 20 GB

Bandwidth High-speed Internet (1Mbps connection)

Software and Third-party tools NLTK 2.0, and SentiWordNet 3.0

Python (v2.7 or above) - *(implementation language):*

Python is a general-purpose, interpreted high-level programming language whose design philosophy emphasizes code readability. Its syntax is clear and expressive. Python has a large and comprehensive standard library and more than 25 thousand extension modules. We use python for developing the backend of the test application. This and the other modules implemented are discussed later.

NLTK (3.3) - (language processing modules and validation):

The Natural Language Processing Toolkit (NLTK) [56] is an open source language processing module of human language in python. Created in 2001 as a part of computational linguistics course in the Department of Computer and Information Science at the University of Pennsylvania. NLTK provides inbuilt support for easy-to- use interfaces over 50 lexicon corpora.

SentiWordNet 3.0:

It is important to note that in case of mining the opinion one can use "SentiWordNet" as it is a lexical resource. There are three scores namely "positivity", "negativity", and "objectivity" of sentiment that is getting assigned for every

SentiWordNet for every synset of WordNet. It groups English words into sets of synonyms called "synsets", provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications.

WordNet: WordNet [57,58]

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing. WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings.

Install any of the few packages to import the Twitter APIs: Tweepy, Tkinter, Textblob. Nltk (natural language toolkit), and matplotlib (for plotting the results on graph). On linux OS these can be installed using the 'pip' command.

```
import numpy as np
import pandas as pd
import nltk
import matplotlib.pyplot as plt
import tweepy
import string

%matplotlib inline
```

Twitter Streaming API

For the purpose of using "Twitter Streaming API", there is requirement that one should first get registered at "http://apps.twitter.com". After the registration is done only it is advised to start using the application page. On the app page then the consumer key is obtained along with consumer secret and then one can create an access token under the "Keys and Access Tokens" tab. Add these to a new file called *config.py*:

```
consumer_key = "add_your_consumer_key"
consumer_secret = "add_your_consumer_secret"
access_token = "add_your_access_token"
access_token_secret = "add_your_access_token_secret"


  # create instance of the tweepy tweet stream listener
  listener = TweetStreamListener()

  # set twitter keys/tokens
  auth = OAuthHandler(consumer_key, consumer_secret)
  auth.set_access_token(access_token, access_token_secret)

  # create instance of the tweepy stream
  stream = Stream(auth, listener)
```

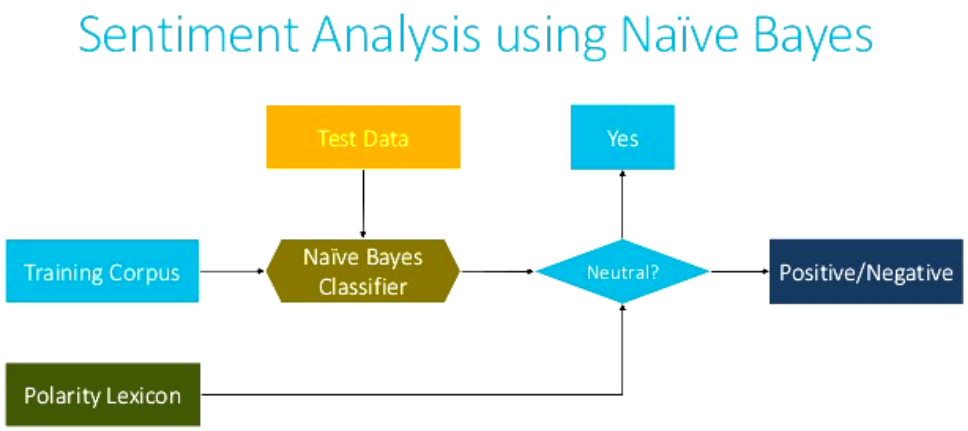

Figure 15: Flowchart of sentiment analysis architecture (using Naive Bayes Classifier)

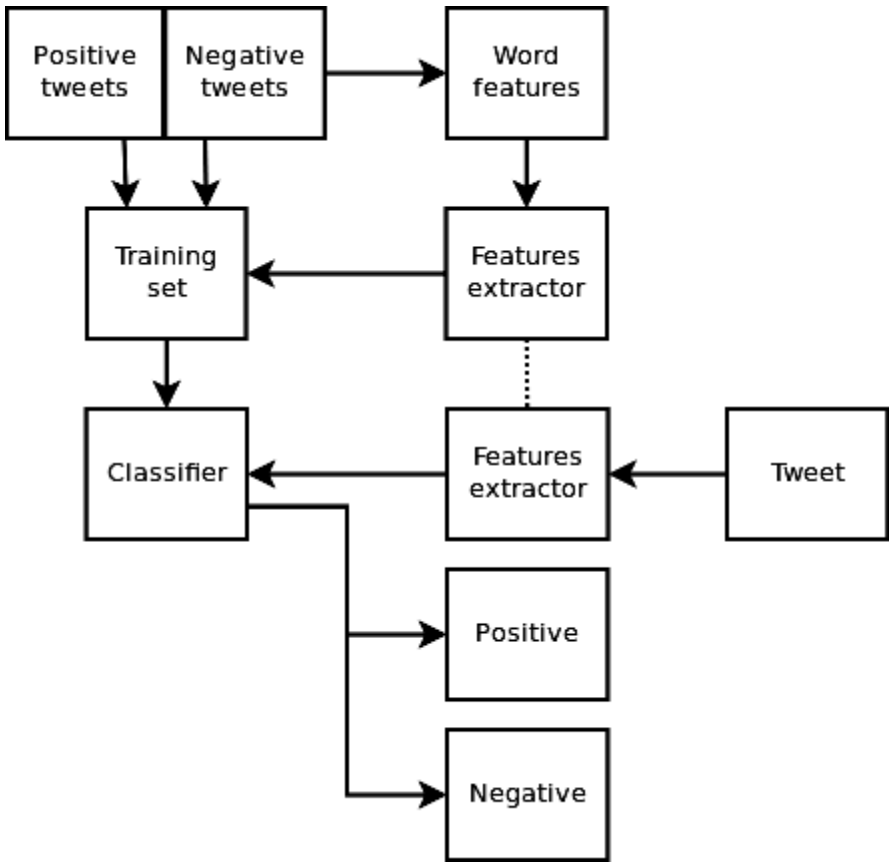

Figure 16: Block diagram for Twitter sentiment analysis

Defining the vocabulary: positive, negative and neutral

```
positive_vocab = {'good','fantastic','terrific', 'astounding', 'outstanding', 'real
negative_vocab = {'bad','worst','terrible', 'useless', 'hate', 'pathetic', 'awful',
neutral_vocab = {'sound','was','home', 'actor', 'know', 'charge', 'world', 'give' }
```

Eliminating words smaller than 2 characters

```
1 tweets = []
2 for (words, sentiment) in pos_tweets + neg_tweets:
3     words_filtered = [e.lower() for e in words.split() if len(e) >= 3]
4     tweets.append((words_filtered, sentiment))
```

Function to calculate the frequency of word occurrence

```
word_features = get_word_features(get_words_in_tweets(tweets))
```

```
def get_words_in_tweets(tweets):
    all_words = []
    for (words, sentiment) in tweets:
      all_words.extend(words)
    return all_words
```

```
def get_word_features(wordlist):
    wordlist = nltk.FreqDist(wordlist)
    word_features = wordlist.keys()
    return word_features
```

```
def train(labeled_featuresets, estimator=ELEProbDist):
   ...

    # Create the P(label) distribution

    label_probdist = estimator(label_freqdist)
   ...
    # Create the P(fval|label, fname) distribution
    feature_probdist = {}
   ...
   return NaiveBayesClassifier(label_probdist, feature_probdist)
```

Results SnapShot

Table 3: Comparison of performances of various analysis methods

	Method	Dataset	Accuracy	Author
Machine Learning	CoTraining	Twitter	82.52%	Liu [59]
	SVM			
	Deep learning	Stanford Sentiment Tree-bank	80.70%	Richard [60]
	SVM	Movie reviews	86.40%	Pang, Lee [61]
	Corpus	Product reviews	74.00%	Turkey
Lexical based	Dictionary	Mechanical Turk (Amazon)	N/A	Taboada [62]
	Ensemble	Amazon	81.00%	Wan, X. [63]
Cross-lingual	Co-Train	Amazon, ITI68	81.30%	Wan, X.
	EWGA	IMDb movie review	>90%	Abbasi, A.
Cross-domain	CLMM	MPQA, N TCIR, ISI	83.02%	Meng [64]
	Active Learning	Book, DVD, Electronics, Kitchen	80% (avg)	Li, S
	Thesaurus			Bollegala [65]
	SFA			Pan S J[15]
Proposed model (Machine	Naïve Bayes	Twitter Movie reviews	92.18%	

```
C:\Python34\Twitter SA\Analyzer\tsanalyze.py
print classifier.show_most_informative_features
Most Informative Features
 idiotic = True          neg : pos    =     37.6 : 1.0
   worst = True          neg : pos    =     32.4 : 1.0
  crying = True          pos : neg    =     24.7 : 1.0
 likeyou = True          neg : pos    =     24.1 : 1.0
    good = True          neg : pos    =     23.4 : 1.0
   hurts = True          neg : pos    =     21.2 : 1.0
   awful = True          neg : pos    =     21.1 : 1.0
   tough = True          neg : pos    =     20.4 : 1.0
terrible = True          neg : pos    =     20.4 : 1.0
   happy = True          neg : pos    =     19.2 : 1.0
  cancel = True          neg : pos    =     19.2 : 1.0
    real = True          neg : pos    =     19.2 : 1.0
pathetic = True          neg : pos    =     18.1 : 1.0
```

Figure 17: Hybrid Naive Bayes feature accuracy on Twitter® dataset

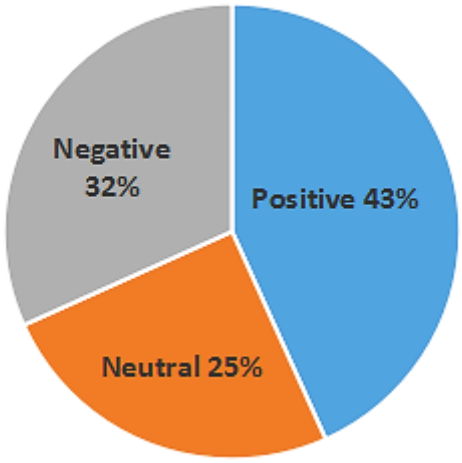

Figure 20: Result of Twitter sentiment analysis

Chapter Six:

CONCLUSION AND FUTURE SCOPE

Present thesis have tried to use some technique of machine learning like "Naive Bayes classifier", "hybrid Naive Bayes classifier with Lexicon based and Natural Language Processing", for the purpose of knowing sentiments inherent in the sentences used in the tweets. A huge data set of 4000 tweets have been used to derive the results which can be banked upon. It has been shown through the research that once we hybridize the prevailing technique of machine learning with technique of lexical analyses in case of classification of sentiment, the probability of fetching accurate results increases manifold.

The consistency that we found in the results was in the range of ≥ 92% - 94% is sufficient to conclude that when we use the hybrid model of machine learning technique with "Naive Bayesian classifier",the results are to good to be applied as model. Further one can use it with confidence in case of similar sentiment analysis application such as "business protocols", "financial sentiment analysis", "customer feedback services", as well as "product surveys" etc.

Chapter Seven:

REFERENCES

BIBLIOGRAPHY

[1] H. Kaur and V. Mangat, "A survey of sentiment analysis techniques.," *In IEEE, 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC),* pp. 921-925, 2017.

[2] J. Sankaranarayanan, H. Samet, Teitle and M. Li, "News in tweets," *In Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems ACM,* pp. 42-51, 2009.

[3] Montoyo, Andrés, P. MartíNez-Barco and A. Balahur, "Subjectivity and sentiment analysis," *An overview of the current state of the area and envisaged developments.,* pp. 675-679, 2012.

[4] G. Hochmuth; , G. Magoulas, B. Lorica and S. Milstein, "Twitter and the micro-messaging revolution," *Communication, connections, and immediacy--140 characters at a time,* 2008.

[5] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis," *In LREC, Twitter as a corpus for sentiment analysis and opinion mining.,* pp. 1320-1326, 2010.

[6] B. Liu, Y. Dai, X. Li, W. Lee and P. Yu, "Building text classifiers using positive and unlabeled examples.," *In Third IEEE International Conference on Data Mining, (ICDM-2003),2003; 179-186,* pp. 179-186, 2003.

[7] R. Jain, K. Chang; , S. Hoi and G. Li, "Micro-blogging sentiment detection by collaborative online learning.," *In IEEE 10th International Conference on Data Mining (ICDM-2010),* pp. 893-898, 2010.

[8] Wawre, S. V., S. N and Deshmukh, "Sentiment classification using machine learning techniques," *International Journal of Science and Research (IJSR),* pp. 819-921, 2016.

[9] L. Huang, R. Bhayani; and A. Go , "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford., 2009.

[10] P. Paroubek and A. Pak, "Twitter as a corpus for sentiment analysis and opinion mining.," *In LREc,* pp. 1320-1326, 2010.

[11] A. Narayanan, I. Arora and A. Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model," *In International Conference on Intelligent Data Engineering and Automated Learning, Springer, Berlin, Heidelberg,* pp. 194-201, 2013.

[12] A. Tumasjan, T. Sprenger, P. Sandner and I. Welpe, "Predicting elections with twitter," *ICWSM,* pp. 178-85, 2010.

[13] J. Bollen, H. Mao and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.," *ICWSM,* pp. 450-3., 2011.

[14] X. Hu, Tang J, H. Gao and Liu, "Unsupervised sentiment analysis with emotional signals," *In Proceedings of the 22nd international conference on World Wide Web, ACM.*, pp. 607-618, 2013.

[15] R. Jose and V. Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using word sense disambiguation.," *In 2015 IEEE International Conference on Control Communication & Computing India,* pp. 638-641, 2015.

[16] T. Apps, "http://www.tweepy.org/," [Online]. Available: http://www.tweepy.org/ (accessed 2018)..

[17] A. Hasan, S. Moin, A. Karim and S. Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts.," *Mathematical and Computational Applications*, p. 11, 2018.

[18] K. Denecke, "Using sentiwordnet for multilingual sentiment analysis.," *In IEEE 24th International Conference on Data Engineering Workshop, 2008. (ICDEW),* pp. 507-512, 2008.

[19] S. Baccianella, A. Esul and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.," *LREC,* pp. 2200-2204, 2010.

[20] G. Miller, "WordNet: An electronic lexical database.," *MIT press,* 1998.

[21] M. Ibrahim, O. Abdillah, A. Wicaksono and M. Adriani, "Buzzer detection and sentiment analysis for predicting presidential election results in a Twitter nation.," *In*

2015 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 1348-1353, 2015.

[22] J. Fang and B. Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification," *InProceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011),* pp. 94-100, 2011.

[23] L. Zhang, R. Ghosh and M. Dekhil, "Combining lexicon based and learning-based methods for twitter sentiment analysis.," HP Laboratories, 2011.

[24] S. Trinh, L. Nguyen and M. Vo, "Combining Lexicon-Based and Learning-Based Methods for Sentiment Analysis for Product Reviews in Vietnamese Language.," *In International Conference on Computer and Information Science, Springer, Cham,* pp. 57-75, 2017.

[25] V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: a survey of techniques," *arXiv preprint,* p. 1601.06971, 2016.

[26] D. Alessia, F. Ferri, P. Grifon and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation," *International Journal of Computer Applications,* p. 125, 2015.

[27] N. Zainuddin and A. Selamat, "Sentiment analysis using support vector machine.," *In International Conference on Computer, Communications, and Control Technology (I4CT 2014),* pp. 333-337, 2014.

[28] A. Sarlan, C. Nadam and S. Basri, "Twitter sentiment analysis," *n IEEE International Conference on Information Technology and Multimedia (ICIMU 2014),* pp. 212-216, 2014.

[29] R. Parikh and M. Movassate, "Sentiment analysis of user-generated twitter updates using various classification techniques.," CS224N, 2009.

[30] A. Yeole, P. Chavan and M. Nikose, "Opinion mining for emotions determination," *IEEE,* pp. 1-5, 2015.

[31] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval,* pp. 1-35, 2008.

[32] L. YM and L. TY, "Deriving market intelligence from microblogs," *Decision Support Systems,* pp. 206-17, 2013.

[33] D. Bollegala, D. Weir and J. Carroll, "Cross-domain sentiment classification using a sentiment sensitive thesaurus," *IEEE transactions on knowledge and data engineering,* vol. 25, no. 8, pp. 1719-31, 2013.

[34] G. Miler, "WordNet: a lexical database for English.," *Communications of the ACM.,* pp. 39-41, 1995.

[35] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques.," in *In Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 2002.

[36] A. Pak and P. Paoubek, "Twitter based system: Using Twitter for disambiguating sentiment ambiguous adjectives.," *In 2010, Proceedings of the 5th International Workshop on Semantic Evaluation (IWSE), Association for Computational Linguistics.,* pp. 436-439, 2010.

[37] J. Almeida and G. Pappa, "witter population sample bias and its impact on predictive outcomes: A case study on elections.," *In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, 2015.

[38] K. Denecke, "Using sentiwordnet for multilingual sentiment analysis.," *In IEEE 24th International Conference on Data Engineering Workshop, (ICDEW 2008).,* pp. 507-512, 2008.

[39] X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu and H. Wang, "Cross-lingual mixture model for sentiment classification.," in *nProceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 2012.

[40] X. Wan, "A comparative study of cross-lingual sentiment classification.," in *InProceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, IEEE Computer Society.,* 2012.

[41] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-based methods for sentiment analysis.," in *Computational linguistics*, 2011.

[42] J. Wu, J. Chuang, C. Manning, A. Ng, C. Potts, R. Socher and A. Perelygin, "Recursive deep models for semantic compositionality over a sentiment treebank.," in *In Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013.

[43] C. Newton, "Twitter just doubled the character limit for tweets to 280.," *The Verge.*, 2017.

[44] B. Badar, W. Kegelmeyer and P. Chew, "Multilingual sentiment analysis using latent semantic indexing and machine learning.," *In IEEE 11th International Conference on Data Mining Workshops (ICDMW, 2011).*, pp. 45-52, 2011.

[45] '. Z. I, P. Saloun, M. Hruzik and I. Zelinka, "Sentiment analysis, e-bussines and e-learning common issue.," *In IEEE 11th International Conference on Emerging e-Learning Technologies and Applications (ICETA*, pp. 339-343, 2013.

[46] M. Lan, C. Tan, J. SU and Y. LU, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE transactions on pattern analysis and machine intelligence,* pp. 721-35, 2009.

[47] M. Klekovkina and E. Kotelnikov, "The automatic sentiment text classification method based on emotional vocabulary," *Digital libraries: advanced methods and technologies, digital collections (RCDL-2012),* pp. 118-23, 2012.

[48] I. Chetviorkin, P. Braslavskiy and N. Loukachevich, "Sentiment analysis track at ROMIP 2011," *Dialog,* 2012.

[49] S. Jiang, G. Pang, M. Wu and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications,* pp. 1503-9, 2012.

[50] T. Koomsubha and P. Vateekul, "A study of sentiment analysis using deep learning techniques on Thai Twitter data.," *In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), IEEE,,* pp. 1-6, 2016.

[51] N. Joshi and S. Itkat, "A survey on feature level sentiment analysis.," *International Journal of Computer Science and Information Technologies.,* pp. 5422-5, 2014.

[52] A. Maas, R. Daly, P. Pham, D. Huang and A. Ng, "Learning word vectors for sentiment analysis.," in *n Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, Association for Computational Linguistics.*, 2011.

[53] Shinde, Dinkar, Pooja and S. Rathod, "A Comparative Study of Sentiment Analysis Techniques.," *IEEE,* 2018.

[54] M. Devika, C. Sunitha and A. Ganesh, "Sentiment analysis: A comparative study on different approaches.," *Procedia Computer Science.,* pp. 44-9, 2016.

[55] S. Vohra and J. Teraiya, "A comparative study of sentiment analysis techniques," *Journal JIKRCE.,* pp. 313-7, 2013.

[56] F. Luo, C. Li and Z. Cao, "Affective-feature-based sentiment analysis using SVM classifier," *In 2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD),* pp. 276-281, 2016.

[57] M. Kamran, A. Noureen, A. Riaz, M. Ali and Q. Ain, "Sentiment analysis using deep learning techniques: a review," *Int J Adv Comput Sci Appl,* p. 424, 2017.

[58] D. Kang and Y. Park, "Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach.," *Expert Systems with Applications.,* pp. 1041-50, 2014.

[59] H. Rui, Y. Liu and A. Whinston, "Whose and what chatter matters? The effect of tweets on movie sales.," *Decision Support Systems,* pp. 863-70, 2013.

[60] P. Gamallo, M. Garcia and Citius, "A naive-bayes strategy for sentiment analysis on english tweets.," *In Proceedings of the 8th international Workshop on Semantic Evaluation (SemEval 2014),* pp. 171-175, 2014.

[61] A. Tsakalidi, S. Sioutas, N. Nodarakis and A. Kanavos, "Large scale implementations for twitter sentiment classification.," *Algorithms.,* p. 33, 2017.

[62] E. Looper, E. Klein and S. Bird, "Natural language processing with Python: analyzing text with the natural language toolkit.," *O'Reilly Media, Inc.,* 2009.

[63] C. Fellbaum, "A semantic network of English verbs. WordNet: An electronic lexical database.153," *IEEE,* pp. 153-78, 1998.

[64] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.," in *In Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 2004.

[65] H. Shen, X. Cheng, F. Li, F. Li and S. Liu, "Adaptive co-training SVM for sentiment classification on tweets.," in *In Proceedings of the 22nd ACM international conference on Information & Knowledge Management, ACM.*, 2013.