Twitter Sentimental Analysis System

Thesis submitted in partial fulfillment of
the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted By

Upasna Joshi

(851232010)

Under the supervision of:

Dr. Parteek Kumar

Assistant Professor

COMPUTER SCIENCE AND
ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

July 2015

Certificate

I hereby certify that the work which is being presented in the thesis entitled, **"Twitter Sentimental Analysis System"**, in partial fulfilment of the requirements for the award of degree of Master of Engineering in Computer Science and Engineering submitted in Computer Science and Engineering Department of Thapar University Patiala, is an authentic record of my own work carried out under the supervision of **Dr. Parteek Kumar**.

The matter presented in this thesis has not been submitted for the award of any other degree of this or any other university.

Upasna Joshi

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

Dr. Parteek Kumar
Assistant Professor
Computer Science and Engineering Department
Thapar University, Patiala

Countersigned by

Dr. Deepak Garg
Head
Computer Science & Engineering Department
Thapar University
Patiala

Dr. S.S Bhatia
Dean (Academic Affairs)
Thapar University
Patiala

Acknowledgement

The successful completion of this project happens to be possible only due to much needed support and able guidance of my seniors and distinguished faculty. It is opportune time to thank one and all for their kind and whole hearted co-operation and best wishes.

First of all I am very thankful to the God for enabling me to fulfil my dream and to complete it. Then I wish to pay my hearty gratitude to Dr. Parteek Kumar for his invaluable guidance, support and inspiration for submitting thesis work on time. I feel honoured and proud to get the opportunity of working under his distinguished guidance.

With the heartiest regards, I wish to express gratitude from the core of my heart towards

Dr. Deepak Garg, Head Computer science Engineering Department for being a Lamppost and for showing me the right path from the very beginning of my career. I would like to thank **Dr. Ashutosh Mishra**, P.G. Coordinator, Computer Science and Engineering Department, Thapar University for the motivation and inspiration in this endeavour.

Finally I would like to express my regards and heartily gratitude to all the members of faculty of computer science department for inspiring and providing me all the facilities needed to complete my work. I would like to thank to my parents and friends for their motivation and whole hearted support in completing thesis work.

Upasna Joshi

Abstract

Internet has become important part of in everyone's life. It is basically useful for user to share their views and opinion in a very short time .Sentimental analysis means extracting important information from user's views. It extract user's opinion posted on sites and micro blogging sites and web form by the user .Thus the extracted information is helpful in decision making process. Sentimental analysis is not a simple process, it is very difficult to exactly predict the sentimental from the text, but there are many challenges in performing sentimental analysis.

The thesis work that is focused on implementation of system based on sentimental analysis. Features are used to extract the sentimental words and computing frequency of each occurring words. This system uses the twitter posts as raw data. It is used to classify sentiment of each tweet. Then it uses the classification algorithm of machine learning algorithm to classify these words. The polarity of each sentiment word is predicated based on these models.

Table of Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Introduction to Sentiment Analysis (SA)

Sentiment analysis or opinion mining is defined as collecting and analyzing the information based upon the user's feelings, thoughts and reviews. Sentimental analysis is also known as opinion mining. Sentimental Analysis uses the concepts of Natural Language Processing (NLP), statistics, machine learning techniques for extraction of subjective information from source materials.

The term Sentimental Analysis refers to automatic extraction of evaluative text, which helps to produce predictive results. Sentimental Analysis and Opinion Mining are very closely related terms. Sentimental analysis first appeared in (*Nasukawa and Yi, 2003*), and the term *opinion mining* first appeared in (*Dave, Lawrence and Pennock, 2003*). However, the research on *sentiments* and *opinions* appeared earlier (*Das and Chen, 2001; Morinaga et al., 2002; Pang, Lee and Vaithyanathan, 2002; Tong, 2001; Turney, 2002; Wiebe, 2000*).

Sentimental Analysis is used by the social media, to generate "public indicator", which means generating the opinions of people by analyzing the results. In Sentimental Analysis solution of one domain cannot directly apply to another domain. Thus, this process is highly domain centred.

Sentimental Analysis has many applications in a real life. It can be used in every phase of life. On social media it has great demand, to get reviews about any product, movies. The purposed system in this study provides the use of sentimental analysis in social media.

Twitter Sentimental Analysis is one of important application of sentimental analysis. The tweets are short form of message which can be posted by the user on his/her twitter account. Twitter

Sentimental Analysis includes the process of extracting tweets and predicts the polarity of tweets, either it is positive, negative or neutral with the help of analyzing process.

1.2 Need of Sentimental Analysis

In today's life there is huge need of Sentimental Analysis. As it automatically analyze and obtained the desired results. The main reasons for using sentimental analysis are discussed as.

1.2.1 Industry Evolution

There is need of only useful data in industry as compared to repository of information. Thus, SA helpful to extract the important information only needed for industry purpose. Sentimental Analysis can give a great opportunity for industries to provide value to their audience and gain value for themselves. Any industry with business to consumer can get benefit from it either it is retail, restaurants/entertainment, hospitality, travel, mobile customer.

1.2.2 Research Demand

Another reason behind growth of SA is research demand in opinion, evaluation, appraisals and their classification. Current solutions for opinion mining and sentiment analysis are quickly evolving, typically by reducing the amount of human effort needed to classify comments. The research theme is based in long established computer science disciplines, such as Natural Language Processing, Text Mining, Machine Learning and Artificial Intelligence, Automated Content Analysis, and Voting Advise Applications.

1.2.3 Decision Making

Every person stores information on web social media, blogs and various web applications, social websites to get relevant information you need some method which analyzes the data and returns useful results. It is very difficult for each company to conduct a survey on regular basis so there is need for analyzing the data and find the best product based on user's reviews, opinions and advices. The opinions and reviews help the people for decision making also helpful for business and research areas.

1.2.4 Understanding Contextual

As human language is very complex and it is difficult for the machine to understand the human language expressed in slang language, nuances, misspelling and cultural variation. So, there is need of system which makes better understanding among machine and human language.

1.2.5 Internet Marketing

Another reason for increase in demand of sentimental analysis is internet marketing by companies and business organization. They regularly monitor user's opinion about their product, brand or event on social post or blog. Thus, Sentimental Analysis can works as a marketing tool too.

1.3 Levels of Sentimental Analysis

Sentiment analysis can be divided into at three levels which are Document Level, Entity or Abstract level and Sentence level.

1.3.1 Document level Analysis

Document level is use to check the document whether it is positive or negative. This type of analysis also called as *document-level-sentimental-analysis.* This analysis is used to evaluate positive or negative opinion about single product. This is only applied to single entity or product, not applicable for multiple entities.

1.3.2 Sentence level Analysis

Sentence level is used to determine the either the sentence is positive, negative or neutral means no opinion. It is about classification of sentences, distinguishes between subjective and objective information in the sentences.

1.3.3 Abstract or Entity Level Analysis

Instead of looking into document, sentences, paragraphs unlike in document level or sentence level analysis, abstract level directly look into the opinion (positive or negative) and target (of opinion). It is also called as *feature-level-analysis* as features are used to determine the

sentiments of the sentences. There are two types of opinion in aspect level analysis, i.e., regular opinion used for particular entity and comparative opinion used for comparison among various entities. This type of analysis used to express problem in a better way.

The document level and sentence level is very challenging. Aspect level divides the problem into sub-problems so it is also complex to use.

1.4 Applications of Sentiment Analysis

Due to efficiency of sentimental analysis, it is in huge demand. Many businesses are adopting text and sentimental analysis to their process. Thousands of text documents can be processed for sentiments. There are numerous application of sentimental analysis. Some of them are described as under.

1.4.1 Word of Mouth (WOM)

Word of Mouth (WOM) is the process of giving information from one person to another person. It helps people in taking decisions. Word of Mouth gives the information about the reactions, opinions or attitudes of consumers about products, business or services that they share with other persons. Hence, this is where SA comes into play. As the online review sites, blogs, social networking sites provide huge amount of opinions, this helps in making decision-making process easier for us.

1.4.2 Voice of Voters

Every political party spent large amount of money in campaigning their party, to influence voters. But if the politicians know the people opinions, suggestions and reviews they can do more effectively. Thus Sentimental analysis not only helps the political parties but as well as helpful for news analysts. Also, American and British administration has used these techniques.

1.4.3 Online Commerce

There is large number of the ecommerce websites. Most of them have policy to get the feedback from the customers and users. After getting the information from the various areas like users experience about product, quality and service details of the company, features and any

suggestions. These reviews and details are collected by the company and convert the data into geographical form with recent updates many online commerce websites uses these techniques. For Example, *Amazon.com*. Thus, SA can bring a big change in company's history.

1.4.4 Voice of the Market (VOM)

Whenever a product is launched by a company the customers wants to know about the product ratings, reviews and detailed descriptions about it. SA can help in analyzing marketing, advertising and for making new strategies for promoting the product. It provides the customer an opportunity to choose the best among the all.

1.4.5 Brand Reputation Management (BRM)

Sentiment analysis helps in determining how company's brand, product or service is being perceived by community online. Brand Reputation Management is concerned about management of reputation in market. It focuses on product and company rather than customer. So, opportunities are created for organizations to manage and strengthen their brand reputation.

1.4.6 Government

Sentiment Analysis helps administration for various services provided to the public. Fair results can be generated to analyze the positive and negative points of government. Thus SA can helps in many fields like taxation, recruitments, decision making policies, evaluating social strategies. Such techniques provide citizen oriented government model where the priorities and services should be provided according to the citizens. Another interesting problem that can be taken up is to apply this method in a multi-lingual country like India, where generating content in a mixture of languages (e.g. English and Bengali) is a common practice.

1.5 Thesis Outline

This thesis report has been divided into 6 chapters. Chapter 1 includes the introduction to sentiment analysis. It also covers applications and levels of sentiment analysis. Chapter 2 describes different approaches of sentiment analysis like knowledge based, concepts based *etc*. The techniques of sentiment analysis such as machine learning and rule based are included in

this chapter. Chapter 3 presents the problem statement, objectives and methodology for developing sentiment analysis system. In chapter 4, Implementation, include the process used for implementing the purposed system. Chapter 5 includes results of the system. Chapter 6 includes conclusion of the work done in this thesis.

Chapter Summary

Sentimental Analysis is process of evaluating author's feeling or opinion and predicts the result. Sentimental Analysis can be used many aspects of our life. Sentimental Analysis has three levels like document level, sentence level and abstract or entity level. With the reasons for increase need in Sentimental Analysis is great increase in demand of industry evaluation and research area. There are many applications of Sentimental Analysis. It can be used for business purposes, in social media, in industries, government. Twitter Sentimental Analysis is important application of Sentimental Analysis. Twitter Sentimental Analysis is used to find the polarity of tweets.

Chapter 2

Literature Survey

2.1 Approaches for Sentiment Analysis

Various approaches can be used for sentimental analysis depending upon the different perspectives. These approaches are keyword based approach, lexical affinity based approach and statistical methods based and discourse structure as shown in Figure 2.1. The description of these approaches is discussed as follows.

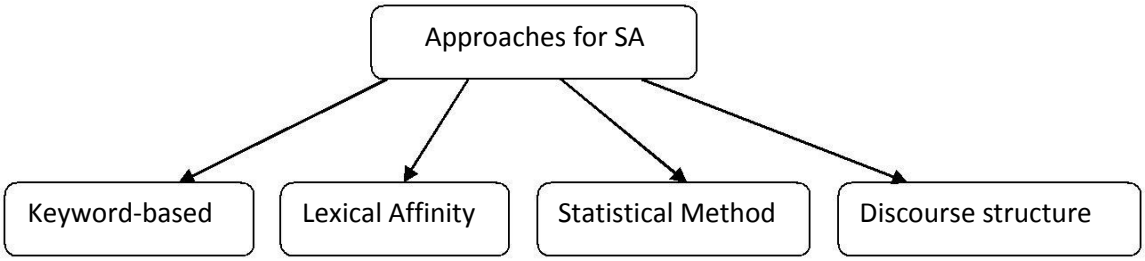

Figure 2.1 Approaches of Sentimental Analysis

2.1.1 Keyword-based Approach

Keyword spotting is the most popular approach used these days due to its economy and simplicity. In this approach text can be classified into ambiguous categories which having affect words such as ‘happy’, ’sad’, ’bored. There are some limitations in this approach which are prescribed below.

Problems in Keyword-based Approach

i. The poor reorganization of text when negation is present and depends on surface features. For example it can simply classify this statement *I am well today* as *positive*. Also recognizes *I*

am sad today as *negative*. But failed to recognize *I am not well today* because it identifies it as *positive* but it is negative statement.

ii. Second problem lies in the sentences which uses strong emotions. Sometimes no effect keyword is used in sentences, and then it is difficult to categorize the statement. For Example, the text *"My husband just filed for divorce and he wants to take custody of my children away from me"* certainly strong emotions, but uses no affect keywords, and therefore, cannot be classified using a keyword spotting approach.

2.1.2 Concept-based Approaches

A large amount of information is present on social media social websites and blogs but when to extract some useful information, is very difficult as this information is unstructured and hence not directly machine-process able concept based analysis can help with this problem. It uses web ontologies and semantic networks to achieve semantic text analysis. Thus, these approaches help the system in extracting the conceptual and affective information from natural language opinions. These approaches don't rely on count of word or any keyword rather it rely on implicit knowledge bases. This analysis level is based on the semantic and affective information associated with natural language opinions and hence able to do feature -based sentiment analysis. Instead of getting individual opinions about a complete item (e.g., Sony Z2), users are generally more interested in comparing different products according to their specific features (e.g., SonyZ2's vs Samsung galaxy S4's touch screen), or even sub-features (e.g., fragility of SonyZ2's v/s Galaxy S4's touch screen). In this, the construction of comprehensive common and common-sense knowledge bases is used for sentiments and polarity detection. So it is essential to properly deconstruct natural language text into sentiments— for example, to appraise the concept "big room" as positive for a hotel review and "big queue" as negative for a bank. Limitation of this approach is it relies of typically knowledge base. Also depend on the depth and breadth of knowledge bases. In fact, it is usually strictly defined and does not allow different concept to be handled, as the inference of semantic and affective features associated with concepts is bounded by the fixed, flat representation.

2.1.3 Lexical Affinity

Lexical affinity approach is somewhat more advanced than keyword-spotting approach. This approach is used to assigns a probabilistic 'affinity' to particular words for a specific emotion rather than simply detecting affect words in the text. For Example, the probability 75% is assigned to word *accident* as *negative*. This approach is better than keyword based approach. This approach has two problems.

(i) First problem is that this approach works on the word-level and can easily be tricked by sentences like given as

I avoided an accident yesterday. ……. *(a)*
I met him by accident ……. *(b)*

In (a) sentence, word "accident" represents in negation form while in sentence (b) the word "accident" specifies another mean of sense.

(ii) Second Problem is that lexical affinities probabilities are depend upon particular domain as specified by the source of the linguistic corpora. Thus, domain independent model cannot be examined.

2.1.4 Discourse Structures

In this approach, discourse relations are used for the classification. For example, in the case of *hotel review* either it is *positive* or *negative* is expressed at the end of the text. Hence this approach is also known as discourse sentimental analysis.

2.2 Techniques for Sentiment Analysis

Sentiment analysis is defined as the process which automatically mines the attitudes, opinions, and emotions from text, speech, and database sources with the help of Natural Language Processing (NLP). Sentiment analysis deals with classifying opinions in text into categories like "positive" or "negative" or "neutral". Techniques used for this approaches are machine learning and Lexicon based approach.

2.2.1 Machine Learning Techniques

Machine learning techniques are defined as the process which involves study of pattern recognization, computational learning theory in artificial intelligence. It is based on construction of algorithm which operates on the model for accepting input and gives some predication data. Thus, such techniques are not static and straight forward. It can be classified into different categories supervised and unsupervised techniques.

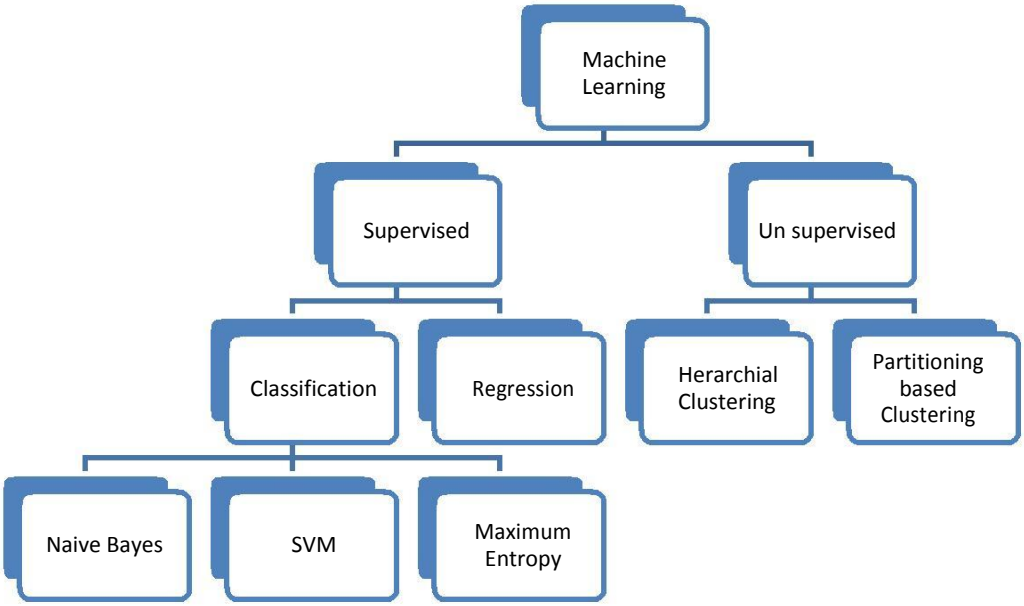

Figure 2.2 Techniques of Machine Learning

2.2.1.1 Supervised Techniques

Supervised machine learning techniques are used to build the prediction model on the basis of set of known data and known responses. This prediction model accepts new set of unknown sample input which generates predicted data. The process includes two phases training of data and another phase is testing of the data.

i. **Training Data**

In this phase the known sample whose desired results are known are trained to obtain classifier model/prediction model.

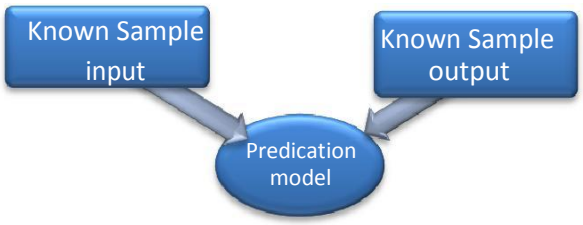

Figure 2.3 Training of data set

ii. Testing of data

Testing of the data help in decision making and predict the result with the help of build model .Figure 2.4 shows the testing procedure.

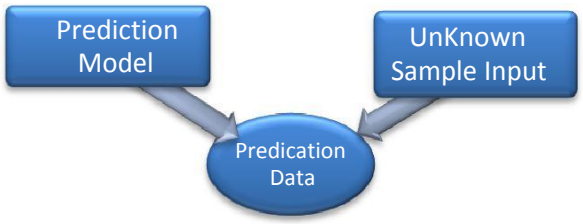

Figure 2.4 Testing of data

Supervised learning can be classified into two main categories as shown in Figure 2.2. First is classification on the known sample data. Classification techniques only accept nominal data set. Second is regression which accepts real number for example miles per second. The classification techniques used many classifiers and algorithms. Mostly used classifiers are Naive Bayes, SVM and Maximum Entropy. They are described as follows.

2.2.1.1.1 Naïve Bayes Classifier

Naïve Bayes classification is based on Bayesian theorem of statistics. A Naive Bayes classifier is a probabilistic model which is based on the Bayes rule with assumption of independance. That in a given class (positive or negative), the words are conditionally independent of each other. This makes the algorithm faster and does not affect its accuracy. In this case, the maximum likelihood probability of a word belonging to a particular class is given by the expression (1).

$$P(x_i|c) = \text{count of } x_i \text{ in a set of a class c / Total number of words in class c} \quad \text{.....(1)}$$

According to the Bayes Rule, the probability of a particular document belonging to class c_i is given by,

$$P(c_i|d) = P(d|c_i) * P(c_i) / P(d) \quad \text{.....(2)}$$

In a given class the words are independent from each other, due to this assumption the model is named as "naïve" and can be expressed as expression (3).

$$P(c_i|d) = (\pi(P(x_i|c_j)) * P(c_j)) / P(d) \quad \text{.....(3)}$$

x_i are the individual words of the document. It is widely used when range of input is large. Due to its simplicity it used for many purposes like online application, text classification, spam filtering and hybrid recommender system. Here is a example of Naïve Bayes theorm.

For example, suppose there is a set of two known sample objects as Red and Green. Our goal is to classify the unknown arrive sample and identify in which category they belongs. Since there are twice Green samples as compared to Red samples. In Naïve Bayes theorem there is a prior probability concept which based on the previous experience. As in this case if a new sample input arrives the prior probability of its Green is greater than being Red.

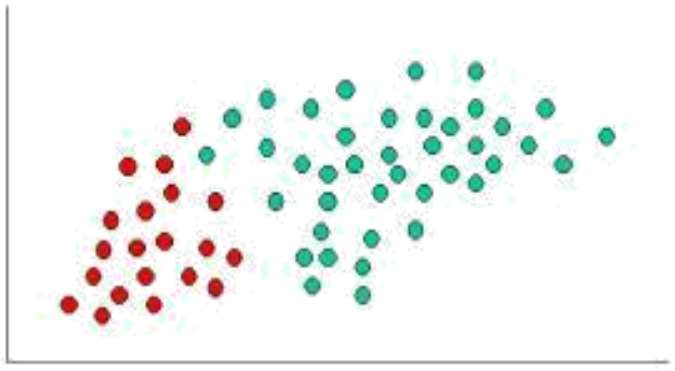

Figure 2.5 Demonstration of Naïve Bayes Classifier

Prior probability for Green $\propto$ *(No. of* Green *objects/Total no. of objects)*

$\propto (40/60)$ ..... (a)

Prior probability for Red $\propto$ *(No. of* Red *objects/Total no. of objects)*

$\propto (20/60)$ ..... (b)

Statement (a) implies prior probability for Green and statement (b) indicates prior probability of Red. Now, if a new object (X) arrives shown as WHITE circle, which can be classified using prior probability. Here these objects are clustered already. The area around the unknown sample can be marked as circle. Here assumption that the number of Green objects is greater as compared to number of Red samples objects inside the encircled area as given in Figure 2.6

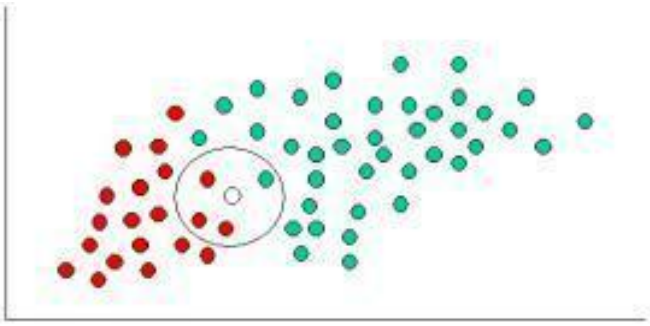

Figure 2.6 Classification of New Object in NBCs

The likelihood of unknown sample to be red and GREEN can be explained with below situation.

Likelihood of X given GREEN ∝ (No. of GREEN in the area of X/ Total no. of GREEN Cases) ∝ (1/40) …….. (c)
Likelihood of X given RED ∝ (No. of RED in the area of X/ Total no. of RED cases) ∝

(3/40) …… (d)

Also prior probabilities of new object (X) can be belongs to GREEN class, shown in statement (c) and (d), but likelihood will belong to RED class. But in this final probability is calculated, by considering two aspects one is prior probability in statement (e) and second is likelihood in statement (f), to generate the posterior probability.

Posterior probability of X being GREEN ∝

Prior probability of GREEN * Likelihood of X given GREEN ...…..(e)

= (4/6) * (1/40) = (1/60)

Posterior probability of X being RED ∝

Prior probability of RED * Likelihood of X given RED …….(f)

= (2/6) * (3/20) = (1/20)

In the end, new sample object X is classified as RED because it has largest posterior probability than others.

2.2.1.1.2 Support Vector Machines

Support vector machines are concept which uses the aspects of decision planes which defines the decision boundaries. A decision planes are used to separate the different classes of objects by any line, curve etc. Such that new object can be easily classified. There are two types of classifiers, *linear classifier* and *hyper plane classifier*.

i. Linear Classifier

It separates the different classes of object by a single line. As given in Figure 2.7 (a). Suppose there are two different classes labelled as Red and Green. If the new sample object is arrived and lies on left side then it is classified as RED but if a new sample object arrives and lies on right side then it is GREEN.

ii. Hyper-Plane Classifier

Mostly classification is not much simple. A classifier that divides the sample set of objects into a set of objects into their respective domains with a curve is called hyper plane classifier where sample objects are separated by any curve or using structure, shown in Figure 2.7 (b).

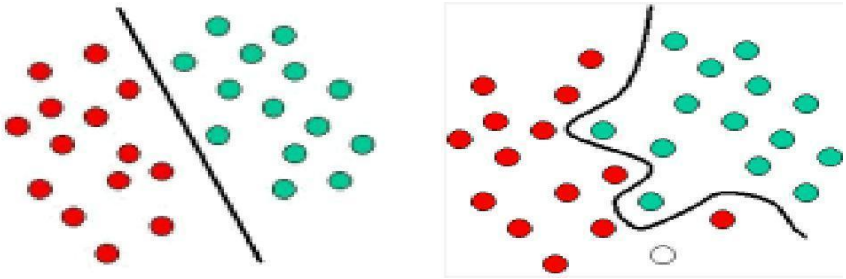

Figure 2.7 (a) Example of linear SVM (b) Example of hyper plane SVM

The basic concept behind Support Vector Machines is shown in Figure2.8. In this figure, original objects are mapped applying a set of mathematical functions known as kernels. This process of reorganizing the objects is known as mapping or transformation. The Figure 2.8 shows that the mapped objects are linearly separable. Thus, find an optimal line rather than constructing the complex curve and to find the optimal line that can divide the Green and the Red objects.

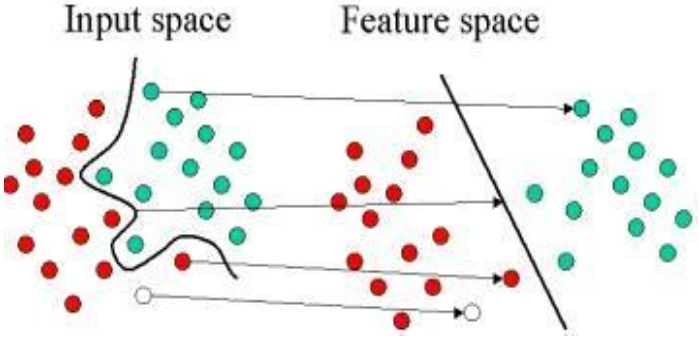

Figure 2.8 Mapping of the objects in SVM

2.2.1.1.3 Maximum Entropy

The drawback of Naïve Bayes classifier is it assumes feature to be independent from each other. For Example, if feature is *best* and the word is *world's best*, it is assumed that both are independent and get multiplied. Thus, the similar words (phrases or idioms) are assumed independent from each other. Maximum Entropy is generally different from Naïve Bayes, it does not depend on independent feature. Also, it is commonly used technique for classification. It is probability distribution technique used for various natural language tasks like text classification, parts of speech, text segmentation and language modelling.

Maximum entropy can handle Boolean, integer and real- valued features. It assumes features to be word level. The outlines for Maximum Entropy Model are prescribed below.

a) Every word w and class c $\in$C, define a joint feature f(w, c)=N where N is the number of times w occurs in a document in a class c.

b) Assigning the weight to each joint feature so as to maximize log- likelihood of the training data.

c) The probability of class c given by a document d and weight λ is shown below in equation(1)

$$P(c|d,\lambda) \stackrel{def}{=} \frac{\exp \sum_i \lambda_i f_i(c,d)}{\sum_{c' \in C} \exp \sum_i \lambda_i f_i(c',d)} \qquad \ldots\ldots\ldots (1)$$

In step (b) there is a search procedure to be involved which is difficult to implement as compared to Naïve Bayes. As in this model the features are correlated to each other thus is prove to be a good model in distributing the weight though.

2.2.1.2 Unsupervised Techniques

Unsupervised learning is another type of machine learning algorithm which is used to extract inferences from database consists of unlabelled data .Unsupervised techniques are hard to examine and understand. There are two types of unsupervised learning

i. First approach is to teach the system not by giving explicit categorizations, but by using some sort of reward system to indicate success.

ii. Second type of approach is the most common unsupervised learning method is cluster analysis, which is used to find data set based on grouping of data. Common clustering algorithms include:

2.2.1.2.1 Hierarchical Clustering

Hierarchical clustering is used to cluster the data in hierarchical manner. It uses two approaches top down approach and bottom up approach.

2.2.1.2.2 Partitioning based Clustering

In this clustering algorithm, data is partitioned. Each item can change the clusters depend on the basis of dissimilarity. K-means clustering algorithm is example of partition based clustering and commonly used.

2.2.2 Feature Extraction for SA

Feature Extraction is the most difficult task in sentimental analysis. Extraction of features will depend on the availability of corpus words. Also feature extraction based on the ontologies so it is domain dependent. There are four categories of features that can be divided .These categories are syntactic, semantic, link-based, and stylistic features discussed as:

2.2.2.1 Syntactic Features

Syntactic features are the mostly used features for sentimental analysis .In this category they can be word n-grams, part of-speech (POS) and punctuation. Also they include phrase patterns which make the use of POS tag n-gram. The phrase pattern where ‘n+aj’ (noun followed by positive adjective) specifies positive statement but ‘n+dj’ (noun followed by negative adjective) is a negative statement.

2.2.2.2 Semantic features

Semantic features include semiautomatic or fully automatic techniques to generate polarity or

affect to the words and phrases. Hatzivassiloglou and McKeown (1997) proposed a semantic orientation (SO) method later extended by Turney (2002) that uses a mutual information calculation to automatically compute the SO score for each word/phrase. Computed score can be formulated as ‘excellent ‘and ‘good’ by getting mutual information between the phrases and the word. Also, the SO approach was later evaluated using latent semantic analysis (Turney and Littman, 2003).

2.2.2.3 Link Based Features

Link-based features use the link analysis to generate the sentiments for web text and documents. Also web text generally has link based structure. It is generally found the web text is linking to each other also having same sentimentent. As the linked based information is used less and having limited use so it is very difficult to analyze the actual use of linked based features.

2.2.2.4 Stylistic Features

Stylistic features include lexical and structural attributes. Here lexical and structural approaches have seen limited usage in sentiment analysis research. These approaches are used for subjectivity and opinion discrimination. It is observed that noticeably higher presence of unique words in subjective texts as compared to objective documents across a Wall Street Journal corpus and noted ‘Apparently, people are creative when they are being opinionated’ used lexical features such as sentence length for sentiment classification of feedback surveys used lexical style markers such as words per message, and words per sentence for affect analysis of web blogs.

Chapter Summary

Sentimental Analysis approaches can be classified as keyword based, sentence based, lexical affinity based approach and discourse structure. Machine learning Techniques are used for sentimental analysis. Machine Learning can be supervised and unsupervised machine learning. Supervised Machine learning take labelled data for testing. This can be Naïve Bayes, SVM, and Maximum Entropy. Unsupervised Learning extract the inferences from database consists of

unlabelled data. Extraction of features depends upon ontologies. Features can be semantic, syntactic, link based and stylistic features.

Chapter 3

Problem Statement

Sentimental analysis is do classification subjective and objective data. The word sentiment means for automatic analysis of evaluative text and mining to produce predictable results. Thus, it helps to extract and classify the relevant information. It uses the various methodologies to classify and simply the opinions. Sentimental analysis is the study of field which analyzes the opinion, sentiments and reviews toward the product, issues, events, places and services.

3.1 Objectives

The main objective or aim of this research work is to use sentimental analysis techniques for analyzing the tweets. To complete this work, following work has been purposed.

i) To analyze the existing techniques and approaches for sentiment analysis.
ii) To extract the twitter posts by user defined parameters using twitter APIs.
iii) To train and test the data using some technique.
iv) To build the classification model.
v) To generate the predicted results based on the unknown data sample.

3.2 Methodology

To achieve the objectives in section 3.1, following methodologies has been used.

i) The Literary Survey consists of study of various approaches to be carried out to perform sentimental analysis. The approaches are supervised machine learning and unsupervised machine learning.

ii) Python language has been used to implement the various approaches for sentimental analysis.

iii) Statement having adjective and nouns, these can be identified and classified according to the classification algorithm and perform the testing.

iv) Use of Machine learning approaches for performing sentimental analysis on tweets.

v) Performing the implementation of system in systematic order involving training nad testing phases.

vi) Apply the classifier and compute the results with their accuracy.

vii) Comparing the results of each classifier.

Chapter 4
Implementation

Sentimental Analysis is defined as the process of evaluating the user's reviews and opinion resulting into sentimental words which are extracted and evaluated by some mechanism giving the predicted output.

4.1 Implementation Details

To implement the purposed system python 3.5 is used with the (Natural language Toolkit) nltk kit .The Implementation process consist of Training the data set and predicting the results.

4.1.1 Role of Python

Python is an interpreter, interactive, object-oriented, high-level programming language. Python plays an important role in sentiment analysis. There are several of modules in python to access Internet and processing Internet protocols. New functions and data types that are implemented in C or C++ are also extended with interpreter of Python. As Python is an extension language so it is suitable for personalized applications.

4.1.2 Role of Natural Language Toolkit

NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python programming language. NLTK has been used successfully as a teaching tool, as an individual study tool, and as a platform for prototyping and building research systems. Natural Language Toolkit (NLTK) is a suite of open source Python libraries for symbolic and statistical natural language processing (NLP). NLTK defines an infrastructure that can be used to build NLP programs in Python. In particular, for building an Automated Sentiment Analysis application, NLTK provides a set of trainable classifiers including a Naïve Bayes Classifier, Maximum Entropy and SVM *etc*.

4.2 Architecture of Twitter Sentimental Analysis

The architecture of the purposed system consists of two parts. First is training the data set and second is predication of unknown data set. The architecture is explained as follow:

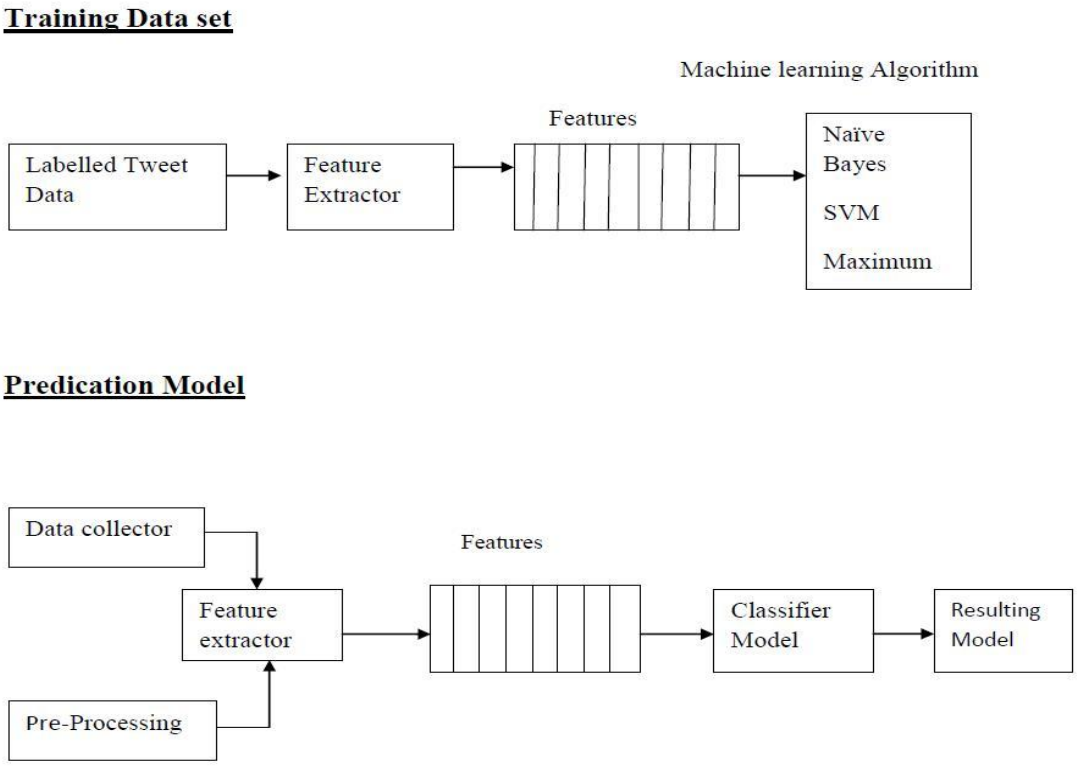

Figure 4.1 Architecture of purposed system

4.2.1 Training Data

Before applying any machine learning algorithm the data set are trained to obtain the desired output. The data set are then goes to predication model to predict the evaluative data. In training data set generally includes following steps.

4.2.1.1 Labelling of data

Twitter is micro-blogging website which allows the users to post the short text message of limited words. The data is first collected form twitter api, tweets are then labelled by the user, either positive or negative.

4.2.1.2 Pre-processing of data

Tweets are different from the words used in textbooks and articles. Due to shortage of words slang language (RT for re tweet, @ for user, # for hash tag, URL's, misspellings) are used by the users to express their views. So, it is very essential to pre-process the information. Pre-processing includes tokenization of the tweets which involves following tasks.

a) Convert the tweets into lower case.

b) Eliminate the URLS's or re-cognize the pattern with the keyword URL.

c) Replace @username with AT_USER and eliminate the # hash tag with exact word. Also remove punctuators extra spaces and repeated words with single letters.

Algorithm 4.1: Pre-processing of tweets Algorithm

Preprocessed (tweets):

tweet= *tweet* to lowercase( ) …… *4.1.1.*

tweet=Convert www. or https?://* to URL* ..... *4.1.2.*

tweet=Convert @user to AT_USER ….. *4.1.3.*

tweet= tweet.remove additional space() ….. *4.1.4.*

tweet=remove the punctuation …… *4.1.5.*

tweet=tweet.strip('\' " ') …… *4.1.6* .

Return(tweet);

Description of Pre-process function:

a) Convert the tweets extracted into lower case as in statement *4.1.1*.

b) Eliminate the URLS's or re-cognize the pattern with the keyword URL described in statement 4.1.2.

c) Replace @username with AT_USER and eliminate the # hash tag with exact word as in statement 4.1.3.

d) Also remove punctuators extra spaces and repeated words with single letters as in statement 4.1.4.

e) Trim the tweets in statement 4.1.6.

4.2.1.3 Feature Vector Extraction

Feature vector are very important for classification models. Every Feature vector helps to determine working of the classifier model. Feature vector help to create the classification model which further helps to predict the unknown data sample. There are various types of feature vectors but in this model, generally used the unigram approach for feature vector. Each tweet words are added to generate the feature vectors. The presence/absence of sentimental word helps to indicate the polarity of the sentences.

Tweets generally consist of various words which are less helpful in result predication. Every polarity type characters so remove the characters which are not meaningful.

a) **Stop Words**: Stop words like a, as, the, above the words which don't indicate the sentiment in the sentences.

b) **Punctuation**: The Punctuation alphabets must consist of punctuation symbols which are removed by extractor.

c**) Repeated Words**: Extractor used to remove repeated words as they don't indicate any sentimental meaning.

Algorithm 4.2: Feature Vector Extraction Algorithm

For row in inptweets

Sentiment=row[0]

Tweet=row[1]

Preprocessedtweet=processTweet(tweet)

Featurevector=getFeaturevector(processedTweet,stopwords)tweets.append(

(featurevector,sentiment));

#end Loop

Description of Feature Vector Extraction

This Algorithm helps to extract the sentimental words from the sentences. Sentimental words are collected.

Algorithm 4.3 Extract Feature Method Algorithm

Def Extract_feature(tweet):

tweet_word=set(tweet)

features={}

for word in featureList: …… 4.3.1

features[‘contains(%s)’%word]=(word in tweet_word) …… 4.3.2

return features

Description of Extract_Feature Function

Extract Feature algorithm used to extract the features from the given data set which is used by the classifiers. In statement 4.3.1 features are extracted one by one by analyzing words in a feature list and in statement 4.3.2 features are collected.

4.2.1.4 Building Classification Model

The classification can be done by supervised machine learning and unsupervised machine learning. In this model we have used supervised machine learning techniques for the classification.

Building the Supervised Model

Supervised learning is very important technique for classification. Mostly used techniques are:

a) Naïve Bayes Classifier

Implementation of Naïve bayes in NLP toolkit

Naïve Bayes classifier is important for classification. Naive Bayes Classifier needs nltk kit to be installed before using this approach. Mostly classifiers are built by training them on training set. To built the naïve bayes classifier in nlp toolkit used ***from*** *nltk.classify.naivebayes* ***import*** *NaiveBayesClassifier.*

Algorithm 4.4 Naïve Bayes Classifier Algorithm

NaiveBayesClassifier(training_set)

testTweet=’Sentimental anlaysis’

ProcessedTestTweet=processTweet(testtweet) …..4.4.1

Print

NBClassifier.classify(extract_features(getFeatureVector(proccessedTestTweet)))
…..4.4.2

Description of Naïve Bayes Classifier

Naïve Bayes classifier accepts the training set. In nlp tool kit naïve bayes classifier can be directly import in program for implementation. In statement 4.4.1 test tweets are processed by the function ProcessedTestTweet. In statement 4.4.2 Naïve Bayes classifier extracts the features of processed tweets and then classifies them.

b) Maximum Entropy Classifier

Maximum Entropy is based on Principle of Maximum Entropy. It fit on one of the model and select the best one which having maximum entropy. This included in Natural Language Processing by Berger and Della Pietra at (1996). Maximum Entropy classifier consider that probability distribution is empirically consistent with the training data if its estimated frequency with which a class and feature vector value co-occur is equal to actual frequency in the data. Many tools are present to implement Maximum Entropy Algorithm.

Implementation of Maximum Entropy in NLP toolkit

Within NLTK, the Maximum Entropy training algorithms support GIS (Generalized Iterative Scaling), IIS (Improved Iterative Scaling), and LM-BFGS. The first two are implemented in NLTK by Python but seems very slow and costs large memory for the same training data.

Algorithm 4.5 Maximum Entropy Classifier Algorithm

```
MaxEntClassifier=nltk.classify.maxent.Maxentclassifier.train(trainingset,'GIS',trace=3,\encodin
g=None,labels=None,sparse=True,gaussian_prior_sigma=0,max_iter=10)

testTweet='Sentimental Analysis'

processedtesttweet=processTweet(testTweet)                    ……4.5.1

print          MaxEntClassifier.classify(extract_feature(getFeatureVector(processedTestTweet)))
……4.5.2
```

Description of Maximum Entropy Algorithm

ProcessedTweet function is used to process the tweets as in statement 4.5.1 after training the data set the classifier accepts the training data in statement 4.5.2 and classify the sentiment words.

c) Support Vector Machine (SVM)

SVM are used for classification, regression. It is used where number of dimensions is greater than number of samples. For the finite data set linear separable plane is used but for the high dimensional data set hyper plane is used.

Implementation of SVM in NLP Toolkit

After training the data set SVM classifier can be used in NLP toolkit. SVM uses *nltk.classify.scikitlearn*. It consists of class nltk.classify.svm.SvmClassifier(trainingset).

Algorithm 4.6 SVM Classifier Algorithm

Classifier= nltk.classify.svm.SvmClassifier.train(train_set) ……4.6.1

for features in test_data_list:

print predict= classifier.classify(extract_features(features)) ….4.6.2

Description of SVM Classifier

SVM classifier is fast efficient as compared to others. It accepts the training data in statement 4.6.1, extract the features and then build the classifier in statement 4.6.2.

4.2.2 Testing of the data

The data set is divided into two parts one set is used to train by classifier and then testinh is done on another set of unknown sample .In testing these classifier are build to generate accuracy and results.

Chapter Summary

As this chapter include the implementation details of each classifier. It describes the method to build the classifier using Natural Language Processing Toolkit. To implement the purposed system python is used to build the classifier model and to obtain the results.

Chapter 5

Results and Discussions

Twitter is micro-blogging site which accepts tweets of only 140 characters. This project helps to analyze the tweets with the help of supervised learning algorithms. The results obtained after applying on the classifiers are as below.

5.1 Training the tweets

Training the data set is required for preparing it for the use of classifiers. In this process it includes the labelling of data, pre-processing the data.

5.1 Labelling the tweets

Before applying the classification model the data extracted using api of twitter. After extraction of the tweets they are labelled manually and list of labelled data is prepared. In Figure 5.1 training data is labelled according to polarity of these sentences as positive and negative.

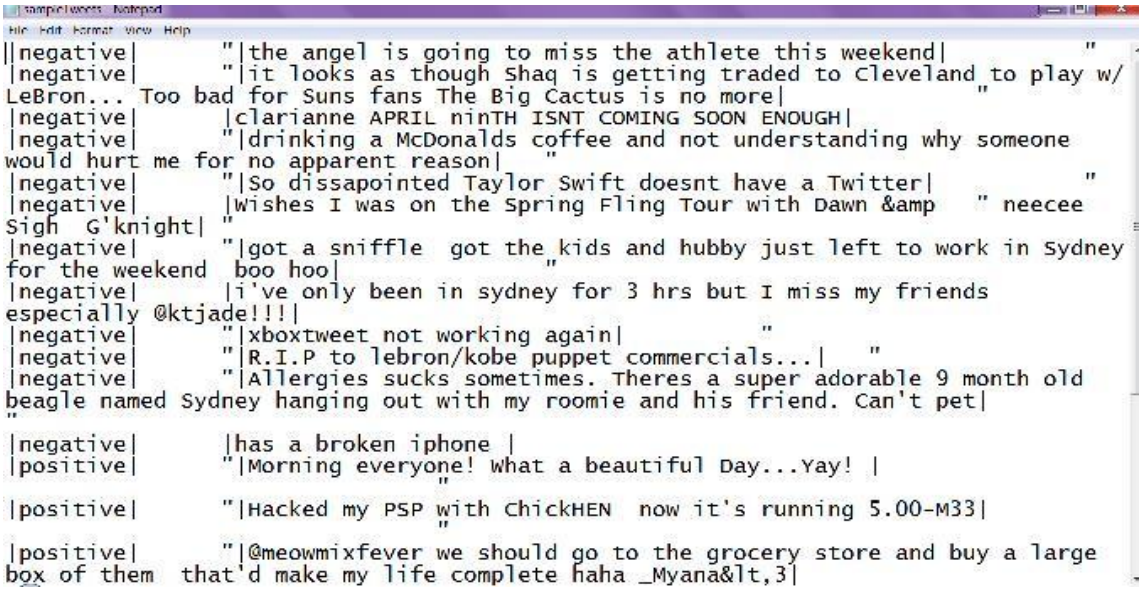

Figure 5.1 Labelling of tweets

5.1.2 Pre-processing the tweets

Twitter tweets consist of words having misspelling, punctuation characters, slang words, URL's.

These words have no meaning in analyzing the sentiment of sentences. Two or more repeated letters in a word are replaced with single word. So, they are required to be removed. Pre processing is process of cleaning the data. In Figure 5.2 consists of pre-processing of tweets

Figure 5.2 Pre-processing of tweets code

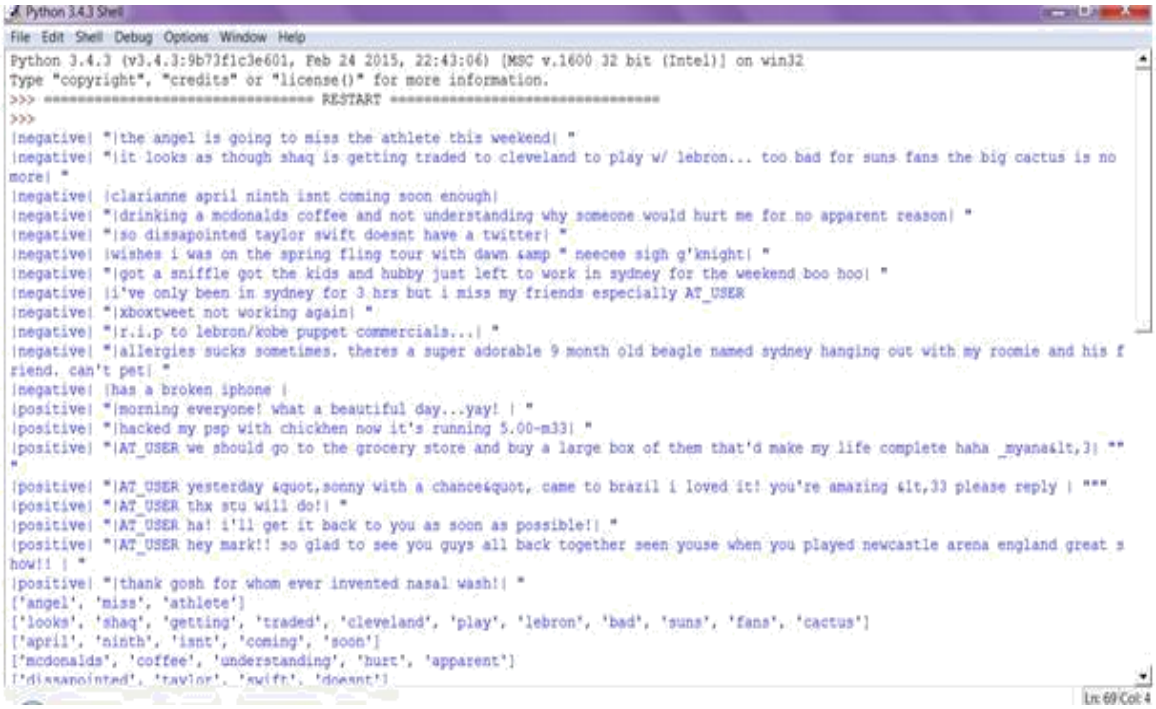

Figure 5.3 Pre-processing Result

5.1.3 Extracting the Feature Vector

For every tweet word, the original tweets are obtained after Pre-processing. This is called feature vector of the concerned tweets. Extraction of tweets is process which takes the feature vector, it check the presence of words in feature vector of the given tweet and present in feature list formed by combining the feature vectors of the tweets. Extracting features process is implemented in python in Figure 5.4. The result of feature extraction is described in Figure 5.5.

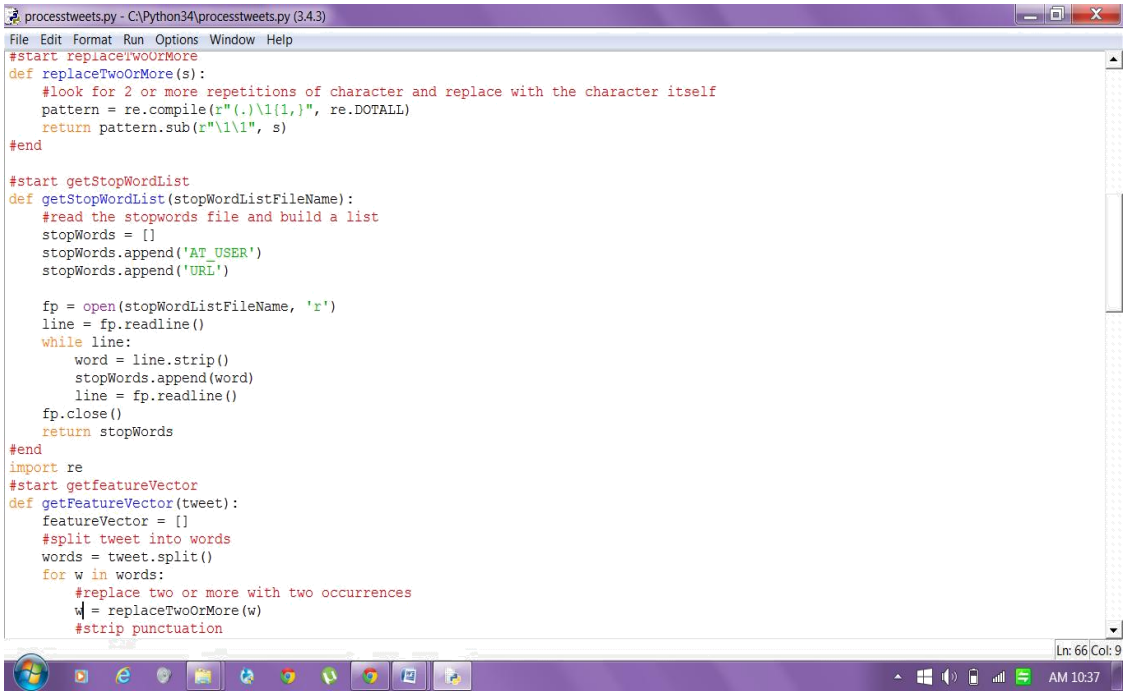

Figure 5.4 Extracting Feature

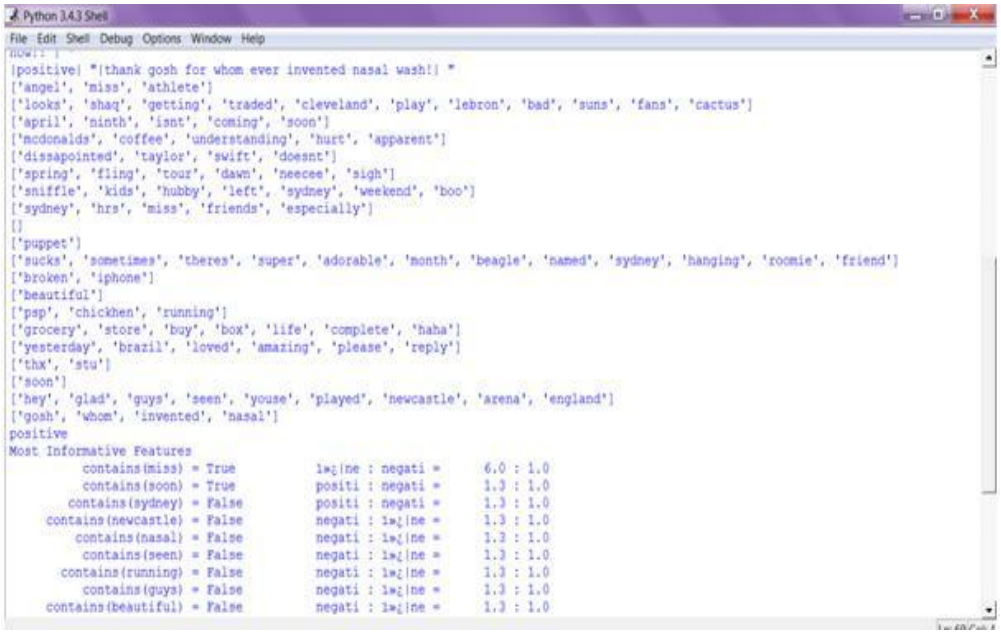

Figure 5.5 Result of Feature Extraction

5.2 Testing the tweets

With the use of build classifier model tweets are tested. Following results shows the testing of tweets.

5.2.1 Building Classifier

Classifiers like Naïve Bayes, Maximum Entropy and SVM are explained in chapter 3, used to test the data to obtain the results. Researchs shows that these classifier shows good range of accuracy and are mostly used over the unstructured data. Depending upon the classifier, the data may need to seprate into training and testing data sets. Classifiers are trained to learn the pattern and in testing classifier use these patterns to evaluate results. In this project there are 2000 tweets which are labelled consists of 1000 positive and 1000 negative tweets.In table 5.1 as shown training and testing record.

Training Set	1600
Testing Set	400

Table 5.1 Training and testing data set

5.2.1.1 Naïve Bayes Classifier

The result of Naïve Bayes is shown in Figure 5.6

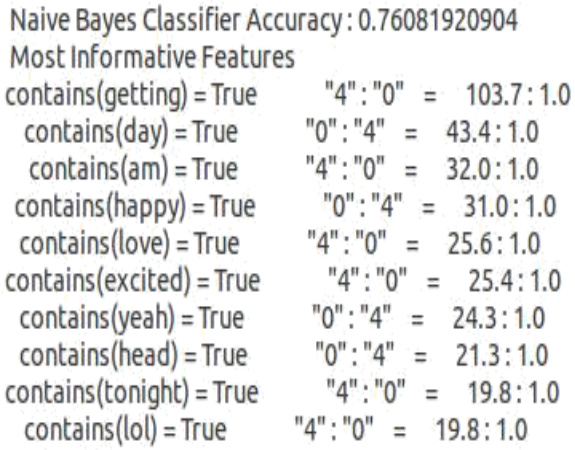

Figure 5.6 Result of Naïve Bayes Classifier

5.3.2 Maximum Entropy

The result of Maximum Entropy Classifier is as shown in Figure 5.7

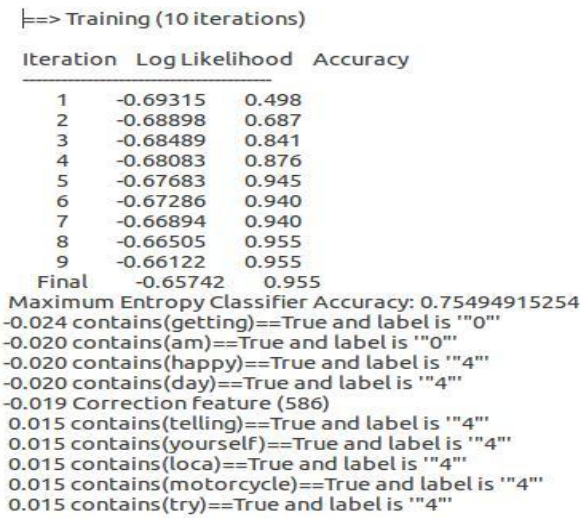

Figure 5.7 Maximum Entropy Classifier

5.3.3 SVM Classifier

SVM Classifier accuracy is approximately 76%.The results shown in figure 5.8.

```
.*
optimization finished, #iter = 5
nu = 0.176245
obj = -2.643822, rho = 0.164343
nSV = 3, nBSV = 0
*
optimization finished, #iter = 1
nu = 0.254149
obj = -2.541494, rho = 0.000000
nSV = 2, nBSV = 0
.*.*
optimization finished, #iter = 6
nu = 0.112431
obj = -1.686866, rho = -0.143522
nSV = 3, nBSV = 0
Total nSV = 4
```

Figure 5.8 SVM Classifier

5.4 Comparison among the classifiers

Comparison among these three classifier can be explained in Table 5.2

Classifier	Accuracy
Naïve Bayes	76%
Maximum Entropy	75%
SVM	77%

Table 5.2 Comparison among classifiers

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

Sentiment analysis is used to identifying people's opinion, attitude and emotional states. The views of the people can be positive or negative. Commonly, parts of speech are used as feature to extract the sentiment of the text. An adjective plays a crucial role in identifying sentiment from parts of speech. Sometimes words having adjective and adverb are used together then it is difficult to identify sentiment and opinion.

To do the sentiment analysis of tweets, the proposed system first extracts the twitter posts from twitter by user. The system can also computes the frequency of each term in tweet. Using machine learning supervised approach help to obtain the results.

The system computes the sentiment of simple English sentences. These sentences are spitted into tokens by tokenization process. Then features extractor used to extract features and results in training set. The classifier is applied on training set to give predictive results on testing set. Also, there are many challenges in sentiment analysis. The purposed system produce accuracy of approximately 75% on tweets.

6.2 Limitations and Future Scope

Some of the limitations of the proposed system and future scope to resolve these limitations are given as follows.

i. The proposed system works for simple sentences only. It can be extended to work for complex sentences also.

ii. Adjective and adverb score files can be improved by adding more sentiment bearing words to it.

iii. The proposed system does not use any parser. So, parser can be embedded into system in future to improve sentiment analysis.

iv. The proposed system is not web based, so it can be extended as web-based application in future.

References

[1] Pang B, Lee L 2008 Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1-13.

[2] Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers May 2012.

[3] Bing Liu, Lei Zhang. *A survey of Opinion Mining and Sentimental Analysis*. Mining Text Data : 415-463.

[4] Karamibekr M, Ghorbani A A 2012 Sentiment analysis of social issues. *Int. Conf. on Social Informatics,* 215-221.

[5] Vohra S, Teraiya J 2013 Applications and Challenges for Sentiment Analysis: A Survey. *Int. Journal of Engineering Research & Technology (IJERT),* 2(2):1-5.

[6] Cambria E, Schuller B, Xia Y, Havasi C 2013 New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems,* 28(2):15-21.

[7] Gebremeskel G 2011 Sentiment Analysis of Twitter posts about news. Master's Thesis, University of Malta.

[8] Annett M, Kondrak G 2008. A comparison of sentiment analysis techniques: Polarizing movie blogs. *Advances in Artificial Intelligence, Springer Berlin Heidelberg,* 25-35.

[9] Ye Q, Zhang Z, Law R 2009 Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications,* 36(3):6527-6535

[10] Sharma A, Dey S 2012 Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis. *Special Issue of Int. Journal of Computer* a*pplications on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA,* 3:15-20.

[11] Witten I H, Frank E 2005 Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann.

[12] Mitchell, T. (1997), Machine Learning. McGraw Hill.

[13] Go A, Bhayani R, Huang L 2009 *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford, 1–12.

[14] Jensen, F. (1996), An Introduction to Bayesian Networks. Springer.

[15] Peng T C, Shih C C 2010 An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs. *IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology*,3: 243-248.

[16] Feldman R. Techniques and applications for sentiment analysis,2013.Commun ACM;56:82–9.

[17] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, March 2000.

[18] An Introduction to Python v2.7.7 2012. [Online] Available: https://docs.python.org/2/.

[19] K. Nigam, J. La®erty, and A. Mccallum. Using maximum entropy for text classi¯cation. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61{67, 1999.

[20] Laurent Luce. Twitter sentiment analysis using python and nltk, 2014. [Online; accessed 9-July-2014]

[21] Christopher D Manning and Hinrich Sch¨utze. Foundations of statistical natural language processing. MIT press, 1999

[22] Niek Sanders. Twitter sentiment corpus, 2014. [Online; accessed 9-July-2014]

[23] Twitter Sentiment Analysis.[Online] Available: http://help.sentiment140.com/for-students/.

[24] How To Build Twitter Sentiment Analyzer. [Online] Available: http://ravikiranj.net/ posts/2012/code/how-build-twitter-sentiment-analyzer/.

[25] NaïveBayesClassifier. [Online] Available: http://en.wikipedia.org/wiki/Naive_Bayes_classifier.

[26] Maximum Entropy Classifier.[Online] Available :http://blog.datumbox.com/machine-learning-tutorial-the-max-entropy-text-classifier/

[27]witter Sentiment Analysis using Python and NLTK.[Online] Available: http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk/

[28]witter Sentiment Corpus.[Online] Available:http://www.sananalytics.com/lab/twitter-sentiment/

[29] Kennedy, Alistair, and Diana Inkpen. "Sentiment classification of movie reviews using

contextual valence shifters." *Computational Intelligence* 22.2 (2006): 110-125.

[30] Das, Sanjiv, and Mike Chen. "Yahoo! for Amazon: Sentiment parsing from small talk on the web." *EFA 2001 Barcelona Meetings*. 2001.

[31] Peng, Fuchun, and Dale Schuurmans. "Combining naive Bayes and n-gram language models for text classification." *Advances in Information Retrieval*. Springer Berlin Heidelberg,2003. 335-350.

[32] Ramamritham K, Bahuman A, Duttagupta S 2006 aAqua: a database-backended multilingual, multimedia community forum. *Proc. Int. Conf. on Management of Data,* Chicago, USA, 784-786

[33] Prabowo R, Thelwall M 2009 Sentiment analysis: A combined approach. *Journal of Informetrics,* 3(2):43-157.

[34] T. Wilson, J. Wiebe, and P. Hormann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language TechnologiesConference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, CA, 2005.