A Novel Approach of Sentiment Detection on Twitter

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

Master of Engineering
in
Software Engineering

Submitted By
Mohit Mertiya
(801431013)

Under the supervision of:
Ms. Ashima Singh
Assistant Professor
CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR UNIVERSITY
PATIALA – 147004

June 2016

Certificate

I hereby certify that the work which is being presented in the thesis entitled, "*A Novel Approach of Sentiment Detection on Twitter*", in partial fulfillment of the requirements for the award of degree of Master of Engineering in *Software 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 *Ms. Ashima Singh* and refers other researcher's work which are duly listed in the reference section.

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

Signature:
(Mohit Mertiya)

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

(Ms. Ashima Singh)
Assistant Professor, CSED

Countersigned by

(Dr. Deepak Garg)
Head
Computer Science and Engineering Department
Thapar University
Patiala

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

Acknowledgement

First of all, I am thankful to God for his blessings and showing me the right direction. With His mercy, it became possible for me to reach so far. I wish to express my deep gratitude to Ms. Ashima Singh, Assistant Professor, Computer Science & Engineering Department for providing her immense help, guidance, simulating suggestions and encouragement all the time. She always provided a motivating and enthusiastic atmosphere to work with. It was a great pleasure to do this thesis under her supervision.

I am also thankful to Dr. Deepak Garg, Head, Computer Science and Engineering Department for his kind help and cooperation. I express my gratitude to all the staff members of Computer Science and Engineering Department for providing infrastructure and encouragement towards research work.

I want to express my appreciation to every person who contributed with either inspirational or actual work to this thesis. Last but not the least I am highly grateful to all my family members for their inspiration and ever encouraging moral support, which enables me to pursue my studies.

Mohit Mertiya
(**801431013)**

Abstract

Twitter has emerged as a platform to express the opinion on various issues. Plenty of approaches like machine learning, information retrieval and Natural Language Processing have been exercised to figure out the sentiment of the tweets. Each of these methods has some benefits and limitations based on the data type used and suitability of data. Most of the research work has been carried out on application of machine learning algorithms applicable for social media sites and further getting the accuracy of the result. However the machine learning algorithms can be integrated with natural language processing algorithms for refining accuracy and context. These refinements tend to increase the accuracy of the result.

In the present thesis, we have purposefully integrated the naive bayes and adjective analysis for finding the polarity of the ambiguous tweets. Experimental outputs have revealed that the overall accuracy of the process is improved using proposed model. Firstly we have applied naive bayes on collected tweets which results in set of truly polarized and falsely polarized tweets. False polarized set has further processed with adjective analysis to determine the polarity of tweets and classify it to be positive or negative. For adjective analysis we have made corpus of adjective negative and positive polarity. We have used movie reviews for our training set as well as test set.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

In this chapter we have discussed about the social media and micro blogging sites, which are becoming the prominent way to express views and opinions. Then we have discussed about the sentiment analysis, which is used to find the sentiments in the posts and texts. We have discussed about the approaches for sentiment analysis and various challenges in the process which affect the performance and accuracy of the process.

1.1 Social media: effects and application

Human being a social animal, interact with each other through their opinions, sentiments etc. Opinions and advice of other has always been an advantage for most of us. Before the introduction of World Wide Web, we used to interact with our friend, relatives and family for piece of advice. In the today's world of modernization, there are numerous ways to share your feelings and opinions. Micro blogging and social networking sites is one of the platforms for communicating with family and friends. It also allows strangers with same interest to connect and share political views, activities and news.

Issue of being at distance is no longer an obstruction for lack of communication. Social sites can be of different forms like blogs, forums, micro blogging, photo sharing, social gaming and video sharing. There are various sites like facebook or twitter which allow to post textual messages. These sites have given us a rostrum to evince our feelings and daily happenings of our life. Twitter, being a micro blogging site allows us to exhibit our emotions through tweets [1]. Figure 1.1 shows different kinds of social media and their applications. It shows that we can post our thoughts through the platforms like faceboook, Google+, Blogs and twitter. All these have great impact on politics, society and culture, some which are as follows:

- Impact on Politics: We have seen the recent trends and posts, where almost each political parties are promoting their leaders and their works through these social media platforms.

- Business Impact: We are shown a number of advertisements, whichever sites we visit. Each company has their promotions and advertisement on these networks.
- Socialization Effect: It has played a great role in reconnecting old friends and making new friends. Social sites like LinkedIn have been used by professionals to find new opportunities and enhance business prospects.

Figure 1.1: Different type of social media

There are some common characteristics of social networking sites which are as follows:

- These are internet based application
- It provides free web space for uploading contents and free web address to identify individuals.
- User created content drives the sites
- User profiles and their data are created and maintained

Figure 1.2 shows different activities that can be done on social networking sites like joining groups, making friends, posting comments, participating in group activities and games etc.

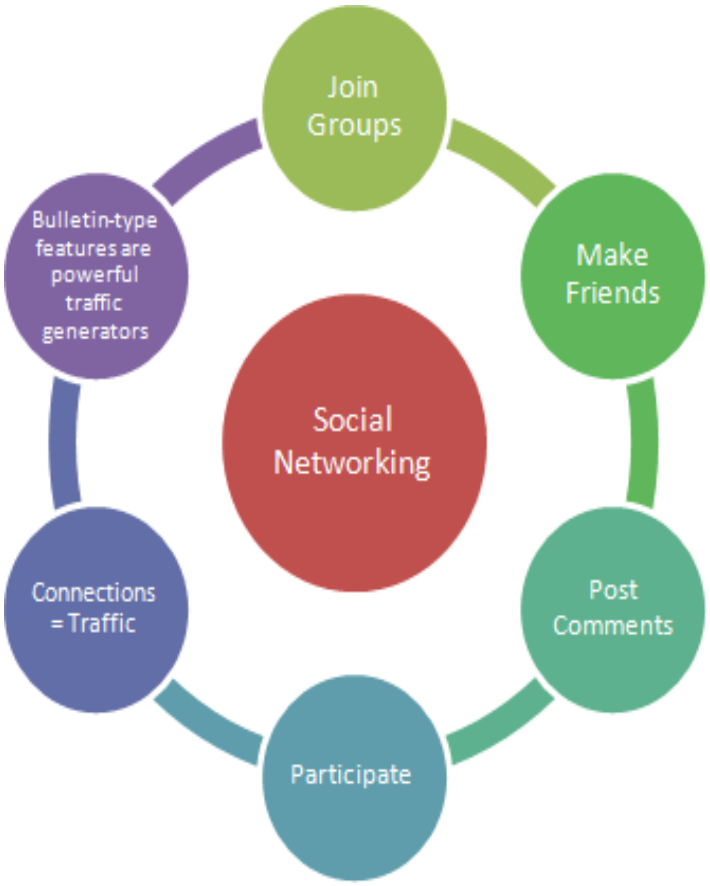

Figure 1.2: various activities of social networking sites

As specified by the statistics millions of users are registered on various sites. Numbers of registered user and services of these micro blogging sites are increasing day by day. According to statistics there are more than 1,590 million facebook users and 320 million twitter users by April 2016. Twitter has approximately 100 million daily active users and has around 2.1 billion search engine queries every day [2]. Over all active social media users are more than 2 billion. These users post their thoughts on the social networking sites. These posts can contain their emotions or can be some factual information. User emotion could be negative or positive. The data from these posts can be used for opinion mining and sentiment analysis, which has wide variety of applications. To predict the emotion in these posts there are various methods available, sentiment analysis is one of those method. Following section discuss about the sentiment analysis and various approaches for the same.

1.2 Sentiment Analysis and Approaches

Sentiment analysis process extracts the emotions from user post. It is the process of discerning the opinion from the text or user posts such as movie review, product review and current trends and then analyzing these opinions to determine the emotion (i.e. negative or positive). Sentiment classification is a salient task in domain of opinion summarizing, reason mining and products comparisons [3]. This helps the companies to review their products and improve accordingly.

In this process we determine that the text written by user is positive or negative. This process is also known as opinion mining. It has wide application areas such as getting product reviews of various customers, social media monitoring, document classification, forecasting and many more. Due to digitization, huge amount of information, reviews and opinions are available on the internet. These opinions give us the perspective of different user groups and useful in variety of ways. It is a great source for extracting views, opinions, attitude, market trends and technologies. All this information is useful for determine the emotions of the user about particular thing. To attain this information, the foremost job is to find the subjective information present in the texts [1].

The scale of work that has been done on twitter is increased in a great way. Twitter sentiment analysis determines the emotion or polarity of the tweet. For analyzing the tweets first and foremost task is to extract the tweet of particular product, topic or any other thing. In our work we have done the polarity classification which analyzes the tweet and classify it as positive or negative. There are various approaches for detecting sentiment in the text as shown in Figure 1.3, broadly categorized into machine learning and lexicon based approach. Lexicon based approach has two types as dictionary based and corpus based. Dictionary based approach is based on collected synonyms and antonyms of words and manual inspection of those words. A dictionary of synonyms and antonyms is constructed, and given text is matched with this dictionary. In corpus based approach a corpus of text posts with various classes created, which is used for training. Natural Language processing techniques analyze the text and generate human readable language.

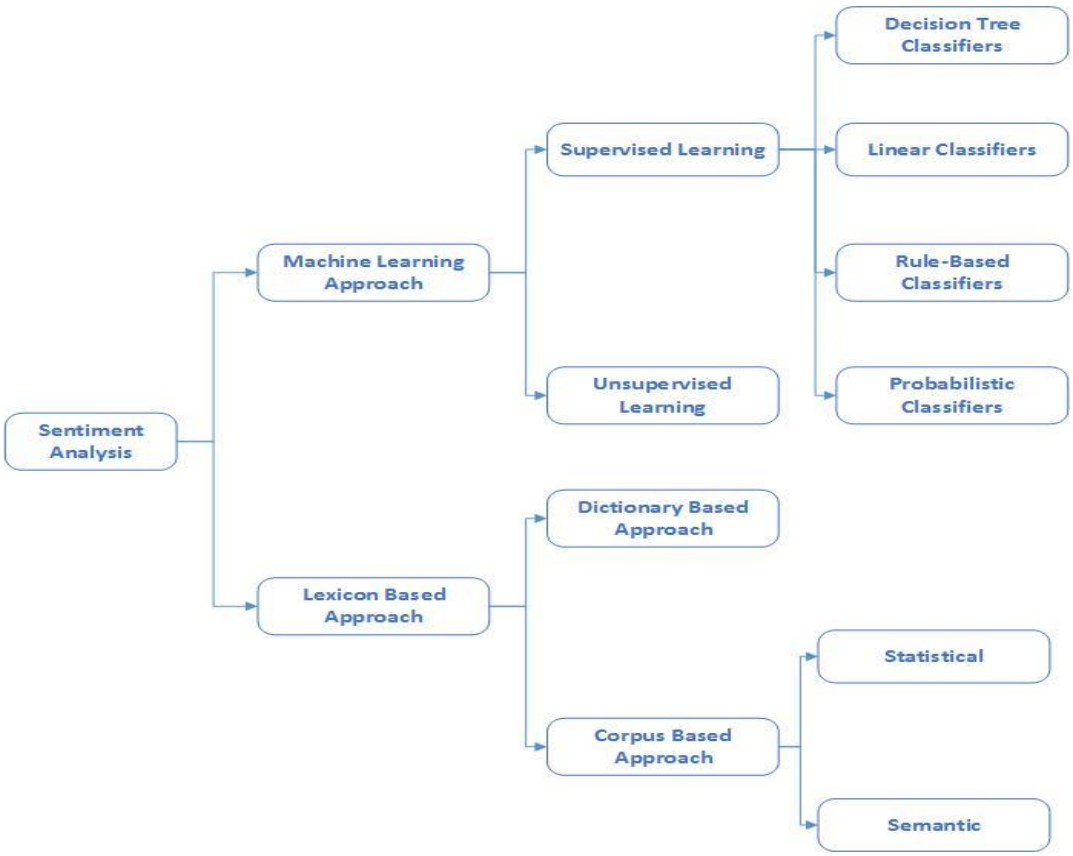

Figure 1.3: Approaches for Sentiment Analysis

Machine learning techniques are classified based on the type of learning as supervised or unsupervised. In supervised learning we train the model using training data, based on the training process model predicts the result as shown in Figure 1.4.

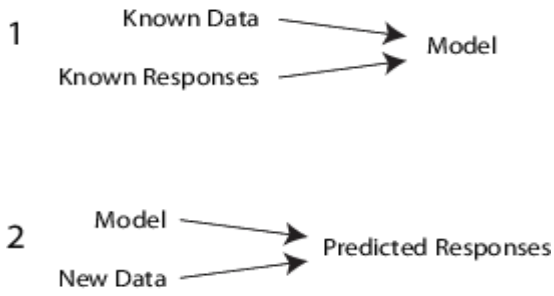

Figure 1.4 Supervised Learning Model

In unsupervised learning model is formed by using structures present in input data. Semi supervised learning combines the characteristics of both supervised and unsupervised.

Figure 1.5 shows the process of sentiment analysis. Detail of each step is explained in later chapters.

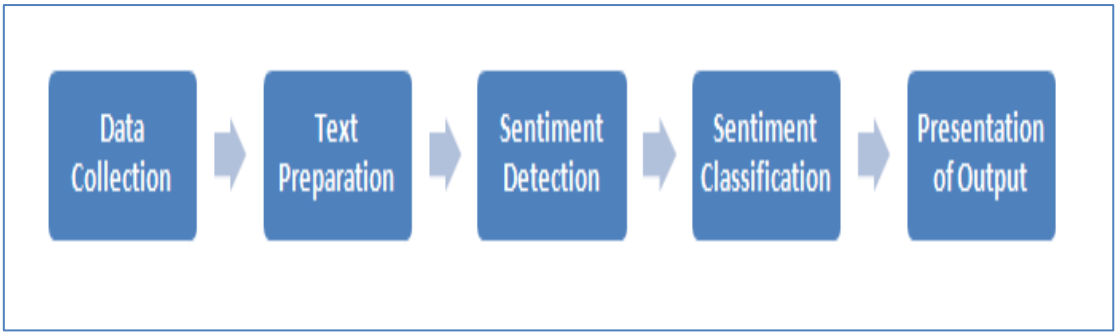

Figure 1.5: Sentiment Analysis Process

The following Table 1.1 shows some of positive and negative tweets collected using #batmanvssuperman and #BritishAirways keyword.

Table 1.1 : Example of positive and negative tweets

Sentiment	Tweet
Negative	Don't fly @BritishAirways. Their customer service and is horrendous.
Positive	Saw #BatmanvsSuperman 6 times now its really worth another watch people
Positive	NO SPOILERS but Batman v Superman is hands down the greatest movie I have ever seen. Every second was perfect. 20/10 in my opinion. #IMAXBVS
Negative	Not even half way through and this movie is such a mess. #bvs not suggested
Negative	#BatmanvsSuperman probably one of the most disappointing films I've seen in a while

As the tweets shows that positive text contains positive keywords such as greatest, perfect etc. Negative tweets contain negative keywords like horrendous, mess, disappointing etc. There are three major steps in process of sentiment analysis:

- Entity Extraction: It is the process of extracting the twitter data automatically using tool or APIs. There are various techniques for the same. The tweets of specific hashtag keywords are extracted.
- Sentiment Existence Identification: This is the process where the presence of sentiment is checked. Selected tweets are analyzed for checking whether they signify some sentiment or they just contain some factual information. It is done by searching for a word or by using a trained classifier like naive bayes, support vector machine etc.
- Sentiment entity attachment: In this step the relation between the sentiment and tweet is detected [4].

These micrcoblogging sites are huge source of information and it is quite easy to say that there is a need of automating the sentiment analysis process as there is too much work involved in processing this information manually. Various approaches are practiced for the automation of this process like machine learning and Natural language processing.

Pak and Paroubek [5] suggested that how to use twitter data as a corpus for analyzing the sentiments and opinion mining. Twitter is one of the most popular social networking platforms that allow users to read and post messages, popularly known as tweets. It can be used through website, mobile app pr SMS. It was created by Jack Dorsey, Noah Glass, Biz Stone and Evan Williams in 2006. It has approximately 500 million users as of 2015. They automated the process of extracting the data and used data for opinion mining and sentiment analysis. The corpus is collected and then Linguistic analysis is performed. Sentiment classifier is implemented using the corpus which find the polarity of the tweet as negative, positive or neutral. Authors have given following reasons for using twitter:

- Twitter is an important source of people's emotions. People tweets on various topics such as their personal life, friends, family, products, stories, news, sports, politics and movies etc.
- It contains a large number of tweets so that collected corpora will be huge. The numbers of tweets are huge in number and increases with a great rate.
- Twitter gives the latest information, news and reviews

- Audience of twitter varies from inexperienced users to experts, student, politicians, sport persons, celebrities, professionals etc. So it covers almost each kind of users which gives us the perspective of all.
- Twitter audience is geographically distributed all over the world. Thus the perspective of each type of user is covered.

1.3 Why sentiment analysis is difficult

Sentiment analysis is very exigent process due to various factors like ambiguity in word sense, contextual dependency, intentional miss-spellings, sarcasm etc. As the tweet size is limited to 140 characters, this makes it more difficult. Testing technique is very critical while validating the process of sentiment analysis. Few of the important variables are data set we are using, size of data set, subject etc. Key challenges in Twitter sentiment analysis are as follows:

- **Anaphora Resolution** - The problem of resolving what a pronoun, or a noun phrase refers to. "We watched the movie and went to dinner; it was awful." What does "It" refer to?
- **Named Entity Recognition** - What is the user actually referring about, e.g. is 300 Spartans a group of Greeks or a movie?
- **Parsing** – determining the subject and object in the sentence, which one does the verb and/or adjective actually refer to?
- **Sarcasm** - If you don't know the author you have no idea whether 'bad' means bad or good.
- **Ambiguity in sentiment words** – it creates conflicts in the analysis process. for example "The movie was terrible" vs. "The movie is terribly good"
- **Negations** – Sentences like "I will never ever say that this movie is worth spending your money on". These negations make it more difficult to judge the polarity.
- **Comparisons in the text** – Sentences like "This product is about as useful as a hole in the head".

- **Quoted and Indirect text** – Sentences of the form as "My friend says the acting in the movie is terrible, but I don't think so".
- **Anything Subtle** – Sentences like "This movie is boring, slow and uninspiring, but it's the acting that saves it".
- **Twitter** - abbreviations, lack of capitals, incorrect spelling and punctuation, poor grammar [6].

The language used in tweets often hinders researcher's potential to analyze sentiment. Jansen et al. [7] carried out the approaches based on term pairs and mutual information, concluding that approximately 80% of the tweets that contained the brands have no sentimental expressions. They analyzed more than 150,000 posts with brand comments, sentiments and opinions. Structure of these posts has been analyzed and various automated methods for classification have been compared. They used a case study to analyze range, content of tweets and frequency.

There are various ways to extract the data or tweets from twitter. Tweets can be extracted automatically by using tools or by using twitter APIs through programs. Each method has its own limitation and drawback. Numerous researchers have encountered challenges while extracting data using twitter APIs. Many foregoing studies extracted tweeter content with Twitter APIs v1.0 but their collected data were restricted in size due to technical difficulties shown by the API and their formulated techniques were not able to perform continuous monitoring of twitter sentiments [8].

Go et al. [9] faced the problem of getting a limit of 100 tweets in response to their query request. Due to this their data sets were limited to set of 100 records every 2 minutes and this lead to non continuous data collection. This discontinuity may lead to inaccurate results. Some of the challenges present in the Twitter API v1.0 were attenuated in v1.1. Twitter4J is another java API which is used for twitter data extraction. It is an unofficial java library that integrates our application with twitter service. It works on jdk version 5 or later, has built in OAuth support, and compatible with Twitter API 1.1. There are four distinct issues associated with analysis of twitter sentiments that need additional exploration:

- Collecting- data: It is the most important part in the process of sentiment analysis. We need to decide the tool or API to use for effective and continuous data extraction.
- Determining sentiment scale to apply to the data: whether to use binary classification by using positive and negative as two classes. Or to use multi class where scale is like strong positive, positive, neutral, negative etc. Key issues while determining the sentiment scale are:
    1. Developing relation between feature and Steps on scale
    2. Developing technique for supervision.
- Feature engineering: It is process of extracting the feature from the documents for the application of machine learning algorithm. Feature is a piece of information that is useful for predicting the output. It reduces the complexity of the analysis.
- Evaluation and classification of Twitter posts: once we are done with feature selection and training the classifier, we evaluate our test set and predict the polarity of the same. [8].

1.4 Organization of Thesis

Chapter 1 discusses the Social Media, Sentiment Analysis Approaches and Challenges faced by researchers.

Chapter 2 discusses various aspects, types and levels of sentiment analysis. It gives the overview and analysis of various approaches being excised for Sentiment Analysis and Opinion Mining.

Chapter 3 states the research gap analysis, formulates the problem and objectives. We have presented the block diagram of our model in this chapter.

Chapter 4 describes the Tools and Techniques used at various stages of our research.

Chapter 5 describes the complete methodology followed in our research. In this chapter we have explained algorithms with example.

Chapter 6 validates the proposed technique using collected Data and compares the results with existing model.

Chapter 7 concludes the work done and stated the future scope of the work

Chapter 2
Literature Review

In this chapter, we discussed about various approaches for sentiment analysis and opinion mining, exercised by various researchers.

As the twitter is briskly emerging communication medium, researchers have worked on numerous methods to observe twitter in real time for various new product reviews, events, stories and reaction of users [8] Researchers in the domain of social network have also worked on the impact of users in the twitter network . Opinion mining and sentiment analysis is vigorous and renowned field in research which is driven by expeditious enhancement in social media and access to huge corpora of information on various issues. It is performed in number of fields of communication. Existing research content delineates two crucial issues: approaches for sentiment analysis process and features applied to depict the unit of writing [10].

Various classes like syntactic, semantic and stylistic have been exercised in writing features. Semantic features contain semi automatically or manually created sentiment lexicons of the positive or negative opinions [11]. For taking the advantage of writing feature in sentiment analysis numerous methods have been recommended like score based and supervised machine learning methods. For analyzing the writing unit score based techniques are used in association with sentiment lexicons. Various machine learning techniques like support vector machine, naive bayes, maximum entropy, neural network are popularly used in sentiment analysis. Few researchers have made their efforts to assimilate the distinctive features of language used in tweets in textual features delineation.

2.1 Aspects of Sentiment Analysis

Most of the study concentrated on semantic properties and lexicons created for other areas. Most of the features of and models are scope oriented and the performance debase when these features are used in other domain [12]. Sentiment detection involves two

aspects as subjectivity classification and sentiment classification. Research work depicted that there is relation between sentiment classification and subjectivity classification. Subjectivity classification can impede polarity classification from contemplating ambiguous and unrelated text. In natural language, subjectivity refers to facet of the language which expresses evaluations and opinions [1]. Subjectivity classification can be defined as follows: In a given document D we have a set of sentences as S={s1,s2,s3,..,sn}. The task of using opinions and other subjectivity (Ss) from sentences to objectively present the factual information (So), where Ss U So=S, is known as subjectivity classification.

Similarity approach is one of the techniques used for the sentence classification. This method analyzes the hypothesis, that for a specifically given topic opinion sentence will be more similar to other opinion sentence than a factual sentence [13]. This process is based on phrases and shared words. It is basically a three step process. First step is to use information retrieval techniques to collect documents of same topic. In second step similarity scores of the sentences present in the document, are calculated. Based on these score an average value is maintained. In the end sentences are assigned to the category for which the average value is highest. Ebert et al. [14] suggested the approach which combines Neural Network and SVM for Sentiment Analysis of Twitter Data. According to authors, users and customers wants the reviews to be short and specific.

2.2 Types and Levels of Sentiment Analysis

Sentiment classification has two types as binary and multiclass classification. Given a set of document as D= {d1,d2,d3...dn}, and a defined categories set C={negative , positive }, binary classification will classify each document di, into label expressed in C. Thus binary classification will have only two values. If the category set is changed to C*= {strong positive, positive, neutral, negative, strong negative}, then classification of the documents di with the labels in C*, then it is called multi-class classification as there are five classes specified [1]. Most of the research work in field of sentiment analysis has emphasized on binary classification of positive vs negative. But having neutral example

is also essential for various reasons. Training the classifier using only positive and negative opinion will limit the accuracy of the process.

Sentiment analysis can be conducted as Natural Language processing task at various levels namely, document level [15][16], sentence level[17][18] and phrase level[19].Pang et al. [16] applied machine learning methods for document classification. Their experimental results suggested that support vector machine performed relatively better than other techniques of classification.

Hu and Liu [20] worked on text summarization and terminology recognition. Their summarization technique was different is a way that they only focused on specific features of the product that customer want to know. They proposed a technique of opinion summarization which works in two steps: firstly extracting features and then opinion orientation identification. Input for their system was product name and output contains reviews summary. Kim and Hovy [21] exercised on different classification models and integrating word and sentence level sentiment. Their algorithm functions in four steps as: choosing the sentence having topic phrase and holder candidate. In second step holder based part of opinion is determined. Then polarity of words having sentiment is calculated by classifiers. Lastly holder's sentiment is created for entire sentence. Wilson et al. [22] developed a system which was capable of assisting Natural language processing applications by imparting subjectivity information in documents.

2.3 Sentiment Analysis Techniques

The most popular techniques for sentiment analysis include machine learning and information retrieval. In machine learning, various algorithms have been developed, like Support Vector Machine (SVM), Naive Bayes. There are three approaches for machine learning as supervised and unsupervised and semi supervised. In Machine Learning classification techniques, we need data sets for training and testing. This data set could be taken collectively or separately.

Classification algorithms use training set to learn several characteristics of text and then classify the test data using this training set. These experimental results connote that

Machine Learning approach gives superior performance than Information Retrieval Methods. Turney and Littman [23] applied unsupervised learning algorithms for classification of reviews. They suggested a method for deducing semantic orientation of a word by using its statistical relationship with set of negative and positive word. Latent semantic analysis and point wise mutual information measures based on co occurrence are used for evaluating the approach. Semantic orientation is effective for review summarization by analyzing highest and lowest semantic orientation. Applications of semantic orientation include automated chat system, software games, and analyzing survey responses.

M Ghiassi et al.[8] worked on supervised learning techniques for applying twitter sentiment lexicons and incorporating distinctive features of language used in tweets. They used Dynamic artificial neural network with uni-gram, bi-gram and tri-gram for feature recognition and classification of tweets. This data driven approach represents the tweet as vector of zeros and ones. Sparse matrix is used for representing the input where one represents that feature is present and zero represents that feature is absent. They have used separate training and test set for evaluation. Two classification methods SVM and Dynamic artificial neural network (DAN2) were used for comparing the effectiveness.

Barbosa and Feng [24] suggested an approach of syntax features of tweets like links, hashtags and punctuations, with features of prior polarity and POS(Part Of Speech) of words and meta information of the composed messages. Their approach is two step process. First step is subjectivity classification of twitter message and then differentiate subjective tweets as negative or positive. They used the data from three websites to train their model and used 1000 manually labelled tweets as test data. Their experimental results revealed that their approach is more productive and robust on noisy and biased data. This improvement was due to the abstract representation of tweets and combining biased data.

Hernández and Sallis [25] proposed an unsupervised technique for analyzing sentiments by following Latent Dirichlet Allocation (LDA). They dealt with the issue of sentiment analysis using probabilistic approach different opinions and sentiments. They segregated

various opinions based on direct relevance to sentiment. They suggested entropy based technique to match word relevance by employing value weighted matrix. Same approach is used for computing document scores. Their method worked on Standard Dirichlet Allocation. They showed that derived reduction has better entropy other models which use complete dataset. Their study involved evaluation with a corpus of 10,000 tweets. The corpus is pre processed and represented with vector space model which uses the TF-IDF (term frequency- inverse document frequency) metric to weight term.

Sharma et al. [26] have worked on the intensities of the words occurring in the sentences using semi-supervised approach. They advised that binary classification doesn't work with fine grained analysis when adjective with common property are compared. Their approach for assigning intensities is based on word vectors rather than on corpus as corpus based technique have the problem of data sparsity. They suggested that text intensity becomes decisive when two sentences are compared for polarity detection and showed the significance of intensity assignment. In such cases they used intensity of various words to determine the polarity of sentences. Words having the same semantic properties can be interchangeably used to enhance or reduce the intensity of sentences. Results depict that this approach performs quite well.

Kouloumpis et al. [27] examined use of linguistic features for analyzing twitter messages. In their study they used three distinct data set as : hashtagged data set from Edinburgh Twitter Corpus for training, Emoticon data set from http://twittersentiment.appspot.com and manually labelled test data from iSieve Corporation. Their study represents that part of speech features are not always be useful in microblogging sites.

Liu et al. [28] presented a model named emoticon smoothed language model for integrating different kinds of data. They suggested training the model using manually labelled data and then using noisy emoticon data for smoothing. They suggested that deficiencies of supervised and distantly supervised techniques can be rectified by using manually labelled and noisy labelled data for training. Their experiments depicted that this approach of integrating different kinds of data can outperform the methods which uses only one of them.

Benamara et al. [29] proposed an approach for sentiment analysis based on adverb adjective combination which uses linguistic analysis. They defined the axioms for scoring the degree of adverbs and adjectives. They proposed three scoring algorithm for adverb adjective combination. They suggested the axiomatic treatment of adverbs and adjectives rather than aggregating the scores of both. Results suggested to have higher precision and recall.

There are some new approaches for opinion extraction which takes opinion consisting of adjectives. There are some shortcomings with this approach like for some domain adjective may not exist or its interpretation could be different for other domain. Harb et al. [30] proposed an approach named Automatic Mining of Opinion Dictionaries abbreviated as AMOD. It is a two step method: firstly extracting learning data set for particular domain and then extracting positive and negative adjective from this learning set. They showed the better performance of the approach with experimental results. Mohammad et al. [31] created a two level SVM classifier, one for detecting sentiment in tweets and one for detecting sentiment of term within tweet. Various features were implemented using surface form and lexical category, in which sentiment lexicons features and ngram features improved the performance.

According to Peng and Park [32] majority of sentiment analysis tools are based on lexicons and machine learning approaches. Building a training set using manually labelled data for large scale application is a challenge in machine learning approach. Sentiment dictionary is used to determine the polarity in lexicon based approach. They represented Constrained Symmetric Nonnegative Factorization Algorithm (CSNMF) which is an automatic sentiment dictionary generation method used to assign polarity score. Their experimental study showed that combination of links from WordNet and corpus performs better in generating dictionary. A lexicon based sentiment analysis is conducted on comments using dictionary to show efficacy of the method.

Prabowo and Thelwall [33] proposed a combination of rule based, supervised learning and machine learning for sentiment analysis, where one classifier helps other to improve the effectiveness. Improvement in F1-measure is shown through this hybrid classification

method using movie reviews, Myspace comments and product review as test data. Barhan and Shakhomirov [34] studied text features for developing method for automatic sentiment analysis of tweets and efficiency was tested thorough polarity determination of tweets. Convolutional neural network was applied by Kim [35] on pre trained vectors for classification of sentences. They proved that CNN works well with hyperparameter tuning and static vectors. Their results shows that supervised pre training of vectors plays an important role in deep learning in Natural language processing. Exchange of information and emotion is done through messaging and posting on social networking site which have the drawback of using informal language that affects the effectiveness of analysis. Affect analysis model [36] was proposed to work on these drawbacks of message which is written in abbreviated or informal style. This model is capable of handling sentences of various complexity levels whether it would be simple, compound or complex. Results of implementing this algorithm showed that it can produce promising results and capable of finding effective information from corpus of blogs and messages. Web text is normally considered to be noisy as it has several problems associated with it at syntactic and lexical level.

Mostafa [37] worked on tweets of some popular brands like Nokia, IBM, DHL, T-Mobile to evaluate the sentiments of the customers. Predefined lexicons having more than 6800 seed adjectives have been used for analysis. Kang et al. [38] proposed senti-lexicon technique for analyzing restaurant reviews. According to authors accuracy of positive classification appears to be roughly 10% higher than negative classification which subsequently decreases the average accuracy when an average value of two classes is taken.

With the help of improved naive bayes algorithm, experiments show that the gap between the positive and negative accuracy was reduced to some extent. Accuracy and recall is improved using senti-lexicon as compared to using svm. Informal writing on social media leads to lower accuracy. Opinion give by users on the blogs can spread with a great rate and negative opinion can be very harmful for the companies. Semantic orientation index and machine learning are the two popular techniques for analyzing sentiments. Semantic

Orientation Indexs have quick response time but have poor performance. Machine learning have better performance but have drawback of long training time [39]. Chen et al. combined the advantages of these methods to propose a new solution based on neural network. Semantic orientation indexes are used as input to neural network and sentiment of the texts extracted from blogs are analyzed. Results of this model show its better performance as compared to traditional approaches by reducing processing time and increasing classification accuracy. Various combinations of algorithms for preprocessing feature selection and classification can also be applied on sentiment analysis process. Al-Daoud [40] proposed a classification model based on four components: feature selection, SVM, Bayesian network and Adaboost. He experiment with seven classifiers for verifying the results and deducted that his model improves the classification accuracy on all dataset. Kontopoulos et al [41] developed an ontology based approach for more efficient sentiment analysis. With their approach tweets are not just classified as sentiment score as in machine learning algorithms by treating each tweet as uniform statement, but assigned a grade for each emotion by breaking each tweet into sets of relevant subjects. This approach has following advantages as: appropriate ontology size, better ontology design and domain specific ontology.

Yessenov and Misailovi [42] presented an empirical study for classifying texts with semantic meaning. They used movie reviews from social network Digg as data set and performed classification by subjectivity and objectivity. They used various techniques for feature extraction as bag-of-word, managing negation, and using WordNet. Accuracy was measured on Decision tree, Maximum Entropy, k-means and Naive Bayes. According to results bag of words performed better andc can be improved by semantic and syntactic information.

Traditional machine learning techniques work well in case of decent match between training and test data within the domain. Read [43] suggested an approach for emoticons labelled training data having match with the domain. He discussed about the dependencies in sentiment classification like topic dependency, domain dependency, and temporal dependency.

Dependency may occur due to the fact that classifiers learn from semantic sentiment of text, not from general sentiments.

Moghaddam and Popowich [44] suggested an approach based on the adjectives present in the text to determine the polarity. Their algorithm first collects the adjectives from the text and computes the frequency of adjectives. Then Naive bayes is trained using positive, negative and neutral classes of adjectives and applied to determine the polarity of the adjectives. In the end classification is done based on weighted average polarity assigned to adjectives. Similarity values of adjectives are calculated using WordNet. Recurrent Neural Network (RNN) are connectionist models that can be used for natural language processing. Irsoy and Cardie [45] applied deep RNN to extract opinion expression at token level. Experimental results deducted that performance of deep and narrow RNNs was better from traditional shallow and wide RNNs. Deep RNNs also performed better than CRF-based models.

Raez et al. [46] worked on polarity classification of tweets, by using vectors of weighted node of WordNet. SentiWordNet has been used to determine final polarity. So this method is unsupervised and domain independent. This method combines random walk algorithm with SentiWordNet polarity scores. Li and Wu [47] used text mining and sentiment analysis on online forums hotspot to detect and forecast. They proposed an algorithm to analyze polarity and obtain value for text. Their algorithm is an unsupervised method which is a combination of k means clustering and support vector machine. First k means is applied on the forum to make the clusters, and then SVM is applied for forecasting and sentiment detection. Experimental outputs reveal that it gives good consistent results. Sentiment detection and forecasting on the hotspot helps companies in various domains as marketing, sales, predicting customer behaviour etc.

Nielsen [48] has examined the ANEW approach termed as Affective norms for English words and other word list to measure the performance for detecting sentiment strength and comparing with list created especially for microblogs. He used manually labelled posts from twitter and by using simple word matching, showed that his method gives better results than ANEW approach. According to him SentiStrength which scores on

word strength, performs better than all other. SentiStrength also includes negation detection, handling spelling variations and emoticons.

Martineau and Finin [49] developed a Delta TFIDF, which is an intuitive general purpose approach to assign weights score on words before classification. They showed the higher accuracy of Delta TFIDF by using SVM. This method generates better results than flat term frequency and TFIDF. TFIDF amplifies values of commonly used terms occurring in documents. Experimental stats shows that it surpasses the performance of raw term count and TFIDF feature weight for subjectivity detection and polarity classification by using uneven distribution of features in different classes.

Denecke [50] introduced multilingual framework for polarity detection of text which is based on lexical resources SentiWordNet. They tested the framework on amazon reviews and compared with n-gram statistical polarity classifier.

In this chapter we have discussed about various approaches and techniques followed by researchers to analyze the text and predict the polarity. We concluded that Machine learning and Natural Language Processing are the most popular approaches. Machine Learning Techniques are Naïve Bayes, Support Vector Machine and Neural Network. Pattern Recognition and Intelligent decisions are the main advantages of these techniques. These algorithms have drawback of learning time and large data requirement for learning process. NLP techniques can analyze text from simple token based representation to rich syntactic representation. Ambiguity in natural language is the main challenge while working with NLP.

Chapter 3
Problem Statement

3.1 Research Gap Analysis

Various approaches and techniques have been followed and applied to detect the sentiments in the texts based on the type of data, and suitability of each technique. Most of the approaches in sentiment analysis are based on applying machine learning algorithm, Information Retrieval and Natural Language Processing. As most of the approaches contained with the application of single algorithm and obtain the sentiments in text, the output is restricted with a set of falsely classified or ambiguous text. This falsely classified text tends to reduce the accuracy of Sentiment Analysis Process. This is due to the fact that some of the tweets can be classified accurately. In case of Machine Learning Techniques the output of the algorithm can be further analyzed and classified which improves the overall performance.

3.2 Problem Formulation and Approach

Sentiment analysis is process of discerning the emotion from the given text and analyzing it to find the polarity. As social networking sites are becoming a rostrum to show our feelings and emotion, the task of classifying text become very crucial. Sentiment analysis has wide areas of application as product reviews, movie reviews, determining market strategy etc.

In our work we proposed a combination of naive bayes and adjective analysis to determine the polarity of the tweets. It has mainly two components as Naive bayes and Adjective analysis component. Firstly the data is collected using the NodeXL tool. This collected data is then pre processed data and fed to naive bayes component which results in polarized and ambiguous set. Our naive bayes is trained with a set of positive and negative reviews stored in text files. The polarized sets are classified as negative and positive set. Set of ambiguous tweets is further given to adjective analysis component where these tweets are classified using polarities of adjectives and adverbs which will be added to positive or negative set. We have made a corpus of negative and positive

adverbs and adjectives. Based on this corpus, ambiguous tweets are classified. Thus the overall accuracy of the sentiment analysis process is improved. The model is depicted in Figure 3.1 showing the components involved in the process of sentiment detection process.

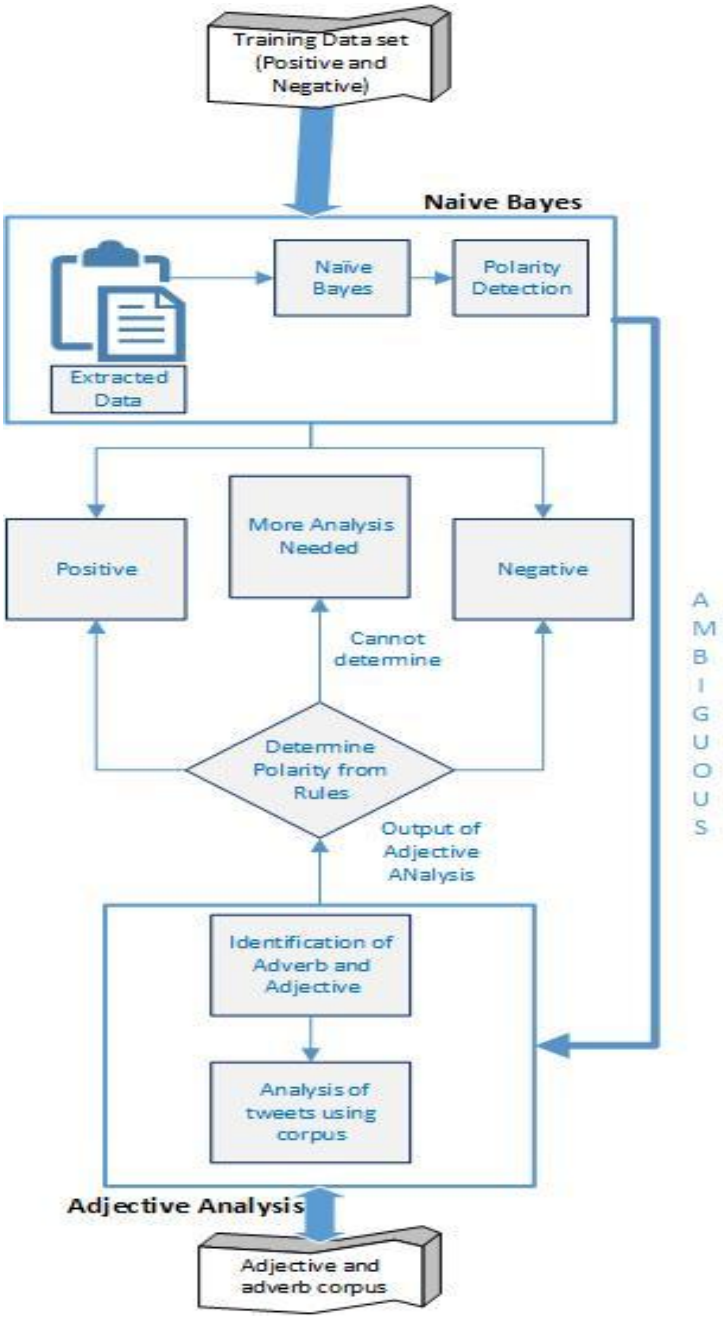

Figure 3.1: Block diagram for the process

3.3 Objectives

i. To study existing approaches in sentiment analysis on social media.
ii. To propose an approach which integrate the Naïve Bayes and Adjective analysis to improve the accuracy of sentiment analysis process.
iii. To validate the approach using the collected tweets by analyzing the results and comparing with previous approaches.

3.4 Scope of the Study

Sentiment analysis has various applications in wide variety of areas. Product Review can be utilized to improve the customer service, recognizing competitors, and enhancing brand reputation. It can also be used to analyze movie reviews. It can also be used to analyze the political issues, decision making, recommendation systems and marketing research.

Chapter 4
Tools and Techniques Used

4.1 NodeXL for Data Collection

Nodexl is an open source tool for extraction and visualisation of social media data. It is basically a template that is embedded into Microsoft Excel on installation. It works with MS Excel 2007 and later versions. It allows us to create network edge list in excel sheet and visualise various graphs. It has two versions as NodeXL Basic and Pro. NodeXL Pro has some extra features over the basic version like advance network metrics, text and sentiment analysis, report generation etc. It allows non programmers to extract essential network statistics and metrics and then visualise the data. It consists of four worksheets: Edge, vertex, Overall Metrics, and Group [51].

Figure 4.1 shows all the features of NodeXL. It shows the Visual properties, Graph, Data, analysis etc.

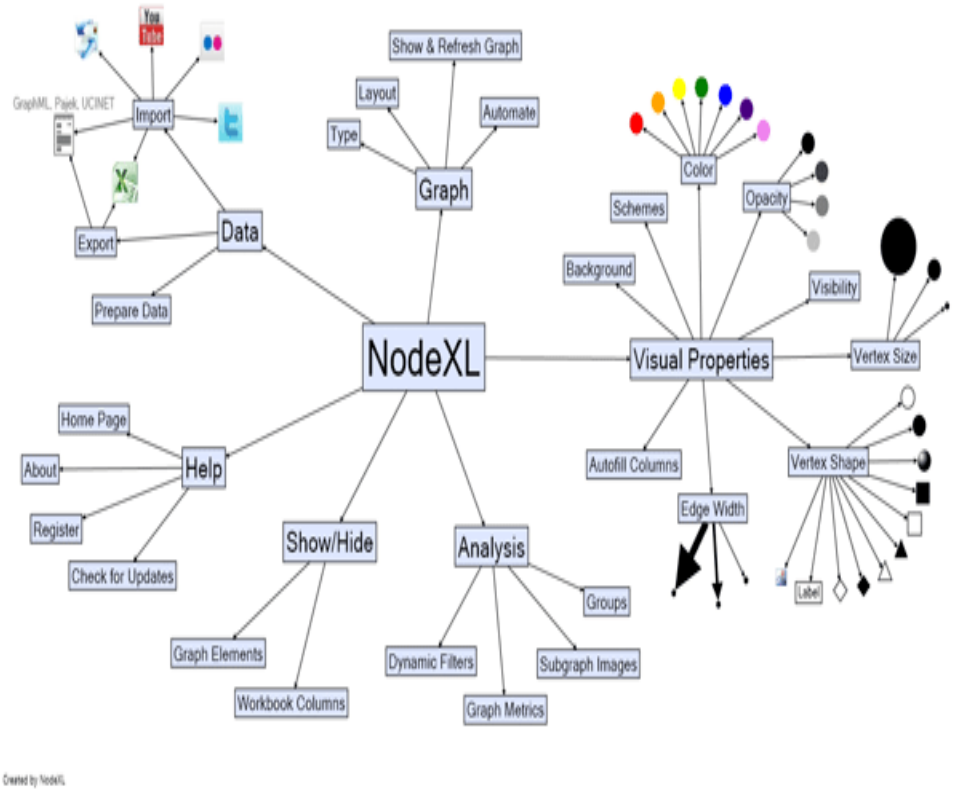

Figure 4.1: NodeXL properties and Features

Following are the features of the NodeXL:

- Flexible Import and Export: it can import and export of data and graphs from other formats.
- Graph Metric Calculation: it provides the facility to calculate network metrics such as degree, centrality, closeness, graph density and many more.
- Connection to social network: it enables us to import data from twitter, flickr, emails and YouTube. It also have various plug-ins which can be used to get data from surveys, hyperlinks and cloud storage.
- Zoom and scale: scaling of graph to reduce clutter.
- Flexible layout: it uses force-directed algorithm for making graph. It allows movement of the vertices with mouse.
- Simply attuned appearance: it enables us to change colour, size, shape of vertices in worksheets.
- Dynamic Filtering and Task Automation: It can perform various repeated tasks automatically.

4.2 R Tool and Libraries

We have used R and R Studio for pre-processing and implementing Naie Bayes. R language is used for graphics and statistical computations. We have installed two libraries in order to execute naive bayes as “RTextTools” and “e1071”. Various functions such as create_matrix(), naivebayes(), predict() and many more, which are used in implementing naive bayes are defined in these libraries.

Chapter 5
Methodology

Figure 5.1 Shows the steps involved in the process starting with data extraction and pre processing. Then naïve bayes is applied on pre processed data which results in classified data sets. Then ambiguous tweets present in the data set are classified using adjective analysis. Complete description of each step is as follows:

5.1 Data Extraction and Preparation

Various techniques are available for collecting the data from the twitter. We have used the NodeXL tool for extraction of the tweets [51]. It includes access to social media which allows us to import the desired data. It can be used to import data from e-mails, Youtube, Twitter and Flickr. Firstly we need to install Nodexl template and embed it into Microsoft excel. After getting Nodexl menu in the Microsoft Excel, we can import the data from the twitter. Then we have to authorize Nodexl by connecting it to twitter. Once the authorization is done, data will be available.

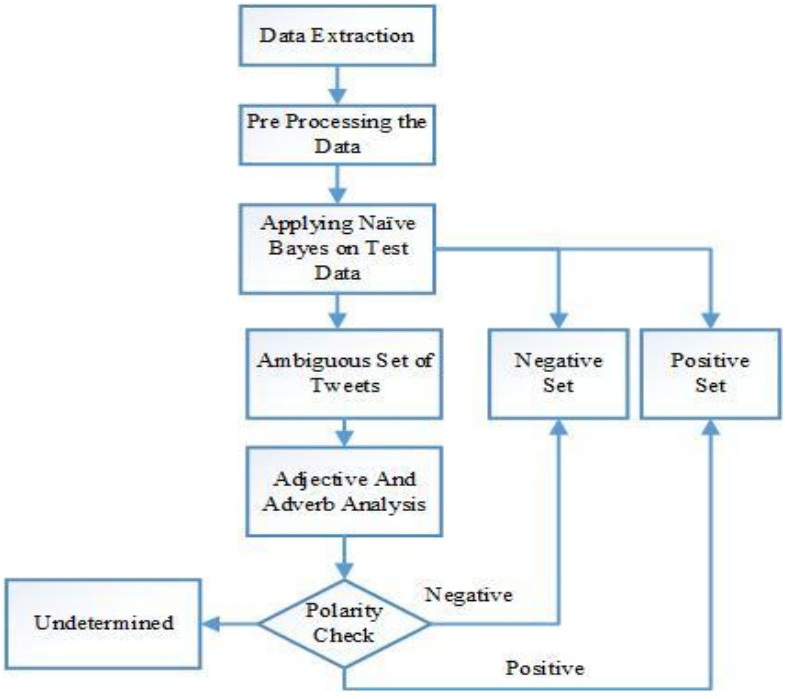

Figure 5.1: Activity diagram for the process

5.2 Pre processing

Twitter posts are very challenging and different from other in several specific properties. Having characters limit of 140, users makes use of abbreviation and similar phrases. Due to this the tasks of tokenizing, lexicon search and part of speech tagging becomes difficult. Apart from these, some special tokens like user mentions, grammatical mistake, spelling mistakes also makes the process difficult. So the extracted data is pre processed to make it more readable. Regular expression can be used to pre process the data.

Mohammad et al. [31] depicted that lexicons are essential for attaining good polarity classification. Incorrect spellings and elongated surface forms of sentiment bearing tokens, like "cooooolll", causes lower coverage of all sentiment lexicons. So in this phase we remove all urls, #hashtags, digits, stop words and @usertags and numbers.

5.3 Naive Bayes Classification

Naive Bayes algorithm is based on simple probabilistic classification, by applying Bayes' theorem with strong independent assumptions. It helps us to calculate the posterior probability as in Eq.1, p(i/j), from p(i), p(j), and p(j/i). This classifier presumes that the effect of the predictor's value on given class is independent of the values of other predictor. This is called as conditional independence [9].

$$p(i/j) = \frac{p(j/i)p(i)}{p(j)} \quad (1)$$

Where i and j are events. *p(i)* and *p(j)* are the probabilities of 'i' and 'j'. *p(i/j)* is the conditional probability of i, given j is true. *p(j/i)* is the probability of j, given i is true.

In our analysis, we have used naive bayes to classify the tweets. Firstly we have created training set to train our classifier. The data collected and pre processed, is used as training set. This training set is fed to the algorithm. Based on this, the polarity of test data is predicted. . Positive training set file contains 1200 positive texts and negative training set contains 1100 negative text. It calculates the likelihood of each word being positive or

negative. When we give some test data, using the probabilities of each word, it finally calculates the polarity of the tweets. After the application of naive bayes we are provided with some False positive and False negative tweets which not correctly classified. To classify these we further applied the adjective analysis.

We have explained the working of naive bayes algorithm with a sample example [52]. Text classification have various application like Reviews classification, mining text, classifing web pages etc. Naive bayes is one of the algorithm used algorithm for text clasiification. It involves following steps:

1. Classify the given document
2. Collecting all the words that are present in training documents
3. Represent each document by vector of words
4. Calculate the required p(c_j) and p($w_k|c_j$) probability terms
    a. For each target value c_j in C do // we have two classes as positive and negative in C
    b. p(c_j)= $\frac{|docs_i|}{|examples|}$ (2)
    c. $Text_j$= single docuent created by concatenating all members of $docs_j$
    d. n = total number of words present in the $Text_j$ (with duplicate words)
    e. for each w_k in vocabulary
        i. n_k=number of times w_k occurs in $Text_j$
        ii. $p(w_k|c_j)$= $\frac{n_k+1}{n+|Vocabulary|}$ (3)

We can write naive bayes conditional independent assumption as:

p(doc|c_j)= $\prod_{i=1}^{length(doc)} P(a_i = w_k|c_j)$ (4)

where p($a_i = w_k|c_j$) is the probability of word in position i is w_k, given c_j

one more assumption p($a_i = w_k|c_j$)= p($a_m = w_k|c_j$) ,for all i and m.

The following example shows the working of naive bayes algorithm. In our example we have taken a small training set and a test set. In out training set we have 5 documents which are classified as negative and positive as shown in Table 5.1.

We have the set of reviews for classification

Table 5.1 : Training Data Document

Text	**Class**
I loved the movie	Positive
I hated the movie	Negative
A great movie, good movie	Positive
Poor acting	Negative
Great acting , a good movie	Positive

Each document is labelled as negative or positive. This training set is used by classifier to train the data and predict the polarity of the given test document.

We represented the documents by vectors of words. We have listed the unique words occurring in the documents. Unique words in the text documents are as follows:

< I, loved, the, movie, hated, a, great, poor, acting, good >

Next step is to convert document into feature sets, where the attributes are possible words, and values are the number of times a word occur in given document as shown in Table 5.2.

Table 5.2 : Count of Words in Training Document

Doc	I	loved	the	movie	hated	a	great	poor	acting	good	Class
1	1	1	1	1							+
2	1		1	1	1						-
3				2	1	1			1		+
4							1	1			-
5				1	1	1		1	1		+

Now calculating the probabilities per outcome (positive or negative)
Firstly we have calculated the probability that document is positive

$p(+)=\frac{3}{5}=0.6$

Now computing probabilities of each word ,given the document is positive we denoted them as:

p(I|+); p(loved|+);p(the|+); p(movie|+); p(hated|+); p(a|+); p(great|+); p(poor|+); p(acting|+); p(good|+)

Now N= number of words in the positive case = 14

Nk= number of times word k occurs in these cases (+)

Vocabulary= list of all distinct and unique words in the training set.

Let $p(w_k|+)= \frac{n_k+1}{n+|Vocabulary|}$ (5)

p(I|+) = $\frac{1+1}{14+10} = 0.083$ p(loved|+)= $\frac{1+1}{14+10}$=0.0833

p(the|+)= $\frac{1+1}{14+10}$=0.0833 p(movie|+)= $\frac{4+1}{14+10}$=0.2083

p(a|+)= $\frac{2+1}{14+10}$=0.125 p(great|+)= $\frac{2+1}{14+10}$=0.125

p(hated|+)=$\frac{0+1}{14+10}$=0.0417 p(poor|+)=$\frac{0+1}{14+10}$=0.0417

p(acting|+)=$\frac{1+1}{14+10}$=0.0833 p(good|+)=$\frac{2+1}{14+10}$=0.125

Now computing the probabilities in case of negative document

Firstly we have calculated the probability that document is positive

p(-)=$\frac{2}{5}$=0.4

Now computing probabilities of each word, given the document is negative we denoted them as:

p(I|-); p(loved|-); p(the|-); p(movie|-); p(hated|-); p(a|-); p(great|-); p(poor|-); p(acting|-); p(good|-)

Now N= number of words in the positive case = 6

Nk= number of times word k occurs in these cases (-)

p(I|-) = $\frac{1+1}{6+10} = 0.125$ p(loved|-)= $\frac{0+1}{6+10}$=0.0625

p(the|-)= $\frac{1+1}{6+10}$=0.125 p(movie|-)= $\frac{1+1}{6+10}$=0.125

p(a|-)= $\frac{0+1}{6+10}$=0.0625 p(great|-)= $\frac{0+1}{6+10}$=0.0625

p(hated|-)=$\frac{1+1}{6+10}$=0.125 p(poor|-)=$\frac{1+1}{6+10}$=0.125

p(acting|-)=$\frac{1+1}{6+10}$=0.125 p(good|-)=$\frac{0+1}{6+10}$=0.0625

Using these probabilities our classifier has been trained. Now we take a new sentence to classify according to:

$$C_{NB} = \underset{c_j \in C}{argmax} \; P(c_j) \prod_{w \in words} P(w|c_j) \quad (6)$$

Where C stands for class

We are taking following sentence as our test document:

"I hated the poor acting"

If C_j= + ; p(+)p(I|+)p(hated|+)p(the|+) p(poor|+) p(acting|+)= $6.03*10^{-7}$

If C_j= - ; p(-)p(I|-)p(hated|-)p(the|-) p(poor|-) p(acting|-)= $1.22*10^{-5}$

The classification will be based on the calculated value of C_{NB} . Calculated values for positive and negative classes are compared and sentence is classified as the one having higher value. Based on the value we deducted that the given test sentence is negative.

After application of Naïve Bayes we are provided with confusion matrix as shown in Figure 5.2, which is used to calculate accuracy. Confusion matrix is used to visualize the performance of classification model. It is a table formed on test data having true and false values where column represent instance of predicted class and row represents instance of actual class. It is also known as error matrix.

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Figure 5.2: Confusion Matrix

5.4 Adjective Analysis

To make our writing more specific and interesting we use adjectives, which describe or modify other words. Adjectives can be used before pronoun, noun or adverbs. Adverbs in the sentence are used to qualify or change the meaning of verb, clause, adjective, adverb or some other word. After the naive bayes is applied we are provided with the results

which contains true positive and true negative, false positive and false negative. False positive and negative do not have their polarity specified and are ambiguous. This data set of ambiguous tweets is further analyzed with adjective and adverbs identification. We have made corpus of adjectives and adverbs which is used to determine the polarity as shown in Figure 5.3. We have worked with following types of sentences: adjectives; verb plus adjectives; adverb plus adjectives.

Sentence consists of adverbs and adjectives. To determine the polarity from these forms of sentences we have formed these rules: two positive adverbs or adjectives will make the final polarity positive; two negative adverbs or adjectives will make the final polarity positive; negative and positive collectively make the final polarity negative.

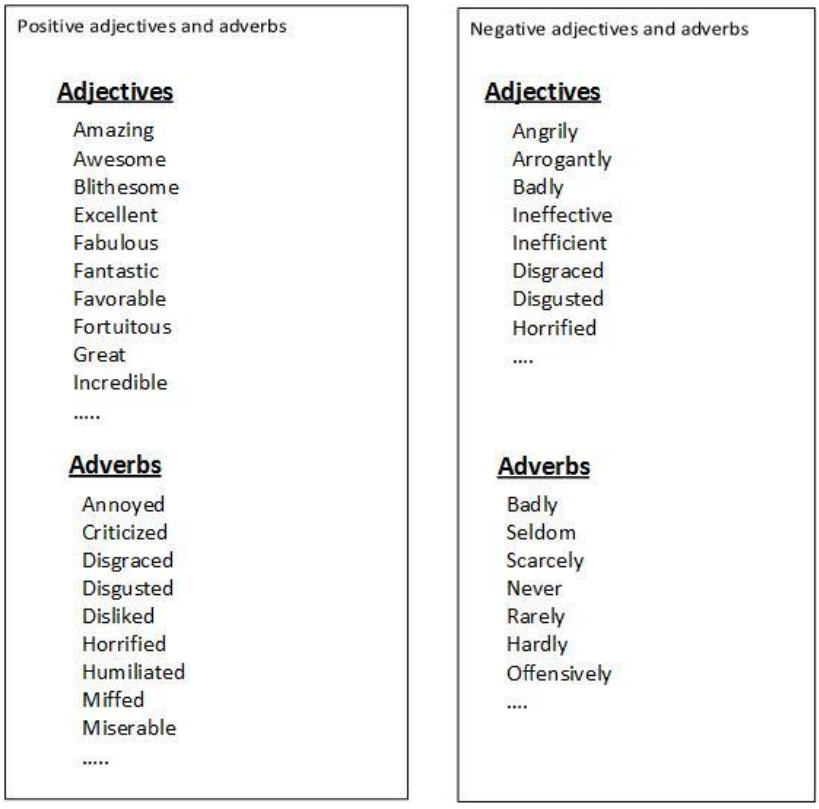

Figure 5.3: List of adjectives and adverbs

From the previous stage output we segregate the ambiguous tweets, which are supplied as input to adjective classification algorithm. The Algorithm for Adjective analysis is as follows:

Algorithm 1: Adjective analysis of ambiguous tweets

Input : T → Set of ambiguous tweets

P→ set of positively polarized tweets

N→ set of negatively polarized tweets

Output: polarized set of tweets

for all t ϵT do

/* checked form adjectives and adverbs table ***/**

if $word_i$ is adjective or adverb

tweet_polarity= polarity[$word_i$]

for every $word_j$ preceding $word_i$ do

if $word_j$ is adjective or adverb

/* polarity is calculated based on specified rules */

tweet_polarity=find_ polarity [$word_i, word_j$]

end if

end for

if tweet_ polarity is negative

N :=N U {t}

else

P:=P U {t}

end if

end for

Algorithm works as follows: once the tweet is selected, adverbs and adjective occurring in the tweet is searched. After this the word which precedes the adjective has analyzed for its nature. Once we are done with finding the adjectives and adverbs, their nature is examined. After that based on specified rules the polarity of the tweet is determined.

Chapter 6
Experimental Setup and Results

R Studio has been used for implementing naïve bayes. We have installed following packages: e1071 and RTextTools. Results obtained at each step are shown. Steps followed in the process are as follows:

6.1 NodeXL Authorization and Data Extraction

We have extracted the data from twitter using nodexl tool. The data is extracted using hashtag keyword search. For this we have to give the keyword for which we have to extract the tweet , then tool will fetch all the related data from the twitter. In this tool we can also specify the number of tweets we want to extract. E.g. Tweets for the keyword #batmanvssuperman. The process of extraction starts with selection of twitter search as shown in Figure 6.1.

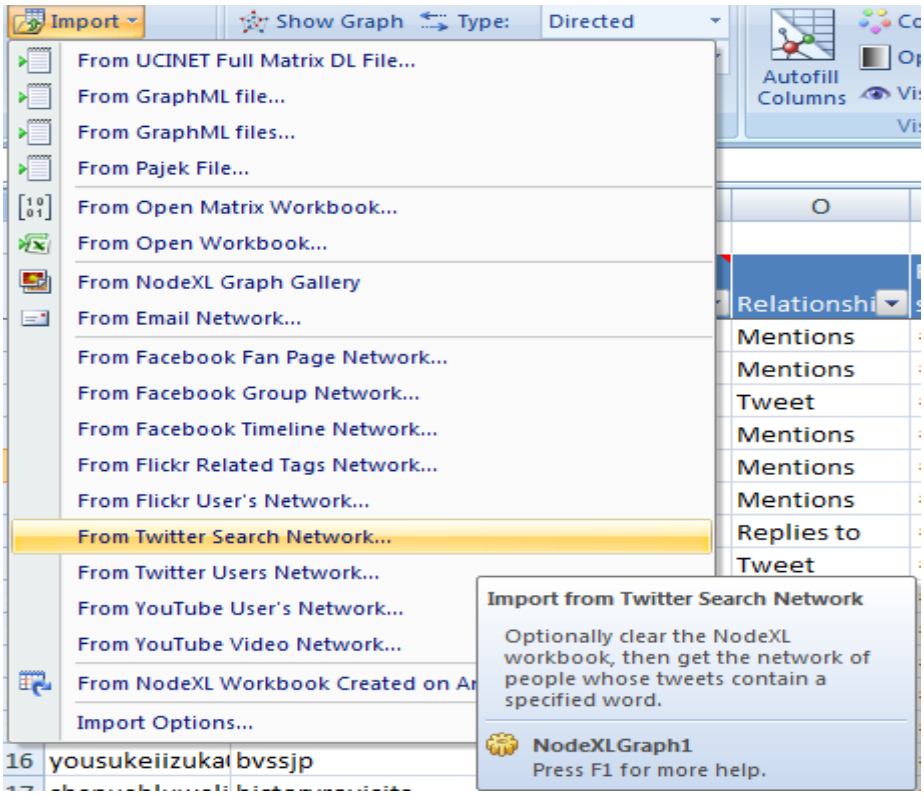

Figure 6.1: Selecting Twitter Network

Figure 6.2 shows the option to authorize and set the limits on number of tweets to be extracted.

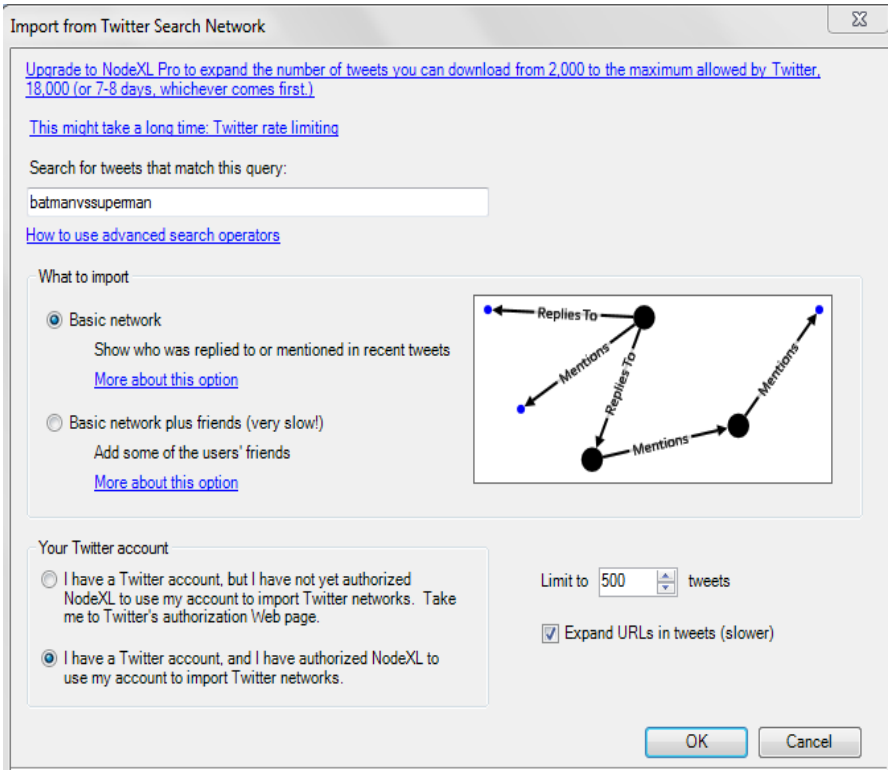

Figure 6.2: Importing data from twitter

We have to authorize the NodeXL through our Twitter account to access the data. By giving the user id and password on twitter account we are provided with a Authorization PIN as shown in Figure 6.3. This PIN is added to NodeXL authorization dialog box, after which we can extract the data and also set the limit for number of tweets to be extracted.

Figure 6.3: Getting PIN from Twitter Account

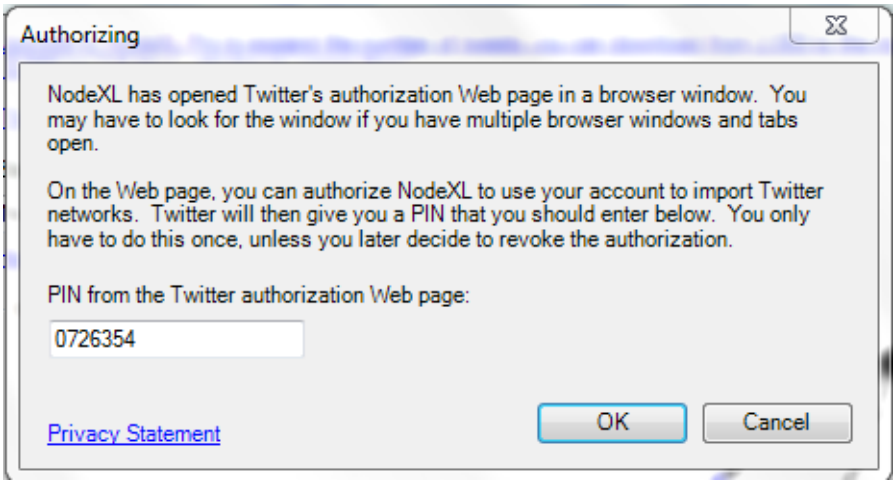

Figure 6.4: Authentication of NodeXL through PIN

Tweet	URLs in Tweet
I saw #BatmanvsSuperman y did Superman die 😳😳	
Todos se burlan de Martha #BatmanvsSuperman https://t.co/pfqC4bhHq5	https://tw
#BatmanVsSuperman al cinema. Scena con Lois Lane sott'acqua, io che dico ALVIIN ALVIIN ALVIIIN. Ridarola	
Notaron en#BatmanvsSuperman Lex Luthor contrata a puros Mexicanos como cocineros y meseros??, Igual	
RT @adamhlavac: Fellow DC fans, it is inevitable that reviewers will compare and contrast #CaptainAmerica	
Ella se llamaba Martha, ella se llamaba así😳 #cantandoando #BatmanvsSuperman	
@FunkoDCLegion #Batmanvssuperman	
RT @seniorH: #BatmanVsSuperman #Batman @Cinemex @WBPictures_Mx https://t.co/FYu9mETYOs	
RT @Stivi_Garca: @Cinemex @WBPictures_Mx #BatmanvsSuperman #QuienGanara #BatmanAmigos, ayúde	
RT @seniorH: #BatmanVsSuperman #Batman @Cinemex @WBPictures_Mx https://t.co/FYu9mETYOs	
RT @seniorH: #BatmanVsSuperman #Batman @Cinemex @WBPictures_Mx https://t.co/FYu9mETYOs	
RT @Stivi_Garca: @Cinemex @WBPictures_Mx #BatmanvsSuperman #QuienGanara #BatmanAmigos, ayúde	
RT @Stivi_Garca: @Cinemex @WBPictures_Mx #BatmanvsSuperman #QuienGanara #BatmanAmigos, ayúde	
RT @adamhlavac: Fellow DC fans, it is inevitable that reviewers will compare and contrast #CaptainAmerica	
RT @_StudioDanielle: Danielle a fait son choix! #BatmanvsSuperman https://t.co/oINF4ckxIx	
RT @adamhlavac: Fellow DC fans, it is inevitable that reviewers will compare and contrast #CaptainAmerica	
RT @adamhlavac: Fellow DC fans, it is inevitable that reviewers will compare and contrast #CaptainAmerica	
#BatmanvsSuperman review is here. #movies #Superheroes @BatmanvSuperman	
Assisti BatmanVsSuperman ontem ❤❤❤❤❤❤❤❤❤❤	
this is a dumb column. I don't have love for #BatmanvsSuperman, but I thought Affleck's Batman	https://tw
RT @adamhlavac: Fellow DC fans, it is inevitable that reviewers will compare and contrast #CaptainAmerica	
Ayer vi #BatmanvsSuperman otra vez y cuando termino, tenía ganas de verla de nuevo.	

Figure 6.5: Extracted Data from NodeXL

Figure 6.5 shows the List of extracted tweets in excel sheet. These tweets contain user references, urls and punctuations.

6.2 Preprocessing using R

In this step collected data is pre processed. We have used R language for the pre processing. Stop words, user references, urls etc are removed from the data. Regular

expressions are used to remove url. Collected tweets are then manually labelled and stored in files as test dataset. We have two data sets: positive and negative.

6.3 Implementation of Naïve Bayes

We have determined the polarity using naive bayes classification algorithm. We have created two training sets: positive and negative. While calling the naive bayes function, we have to specify the size of training and test data in matrix.

```
classifier = naiveBayes(mat[1:2300,], as.factor(sentiment_all[1:2300]))
```

Here the training text ranges from 1 to 2300, as there are 1200 positive review and 1100 negative reviews in our training file. We have created two separate files for positive and negative set as shown in Figure 6.6 and Figure 6.7.

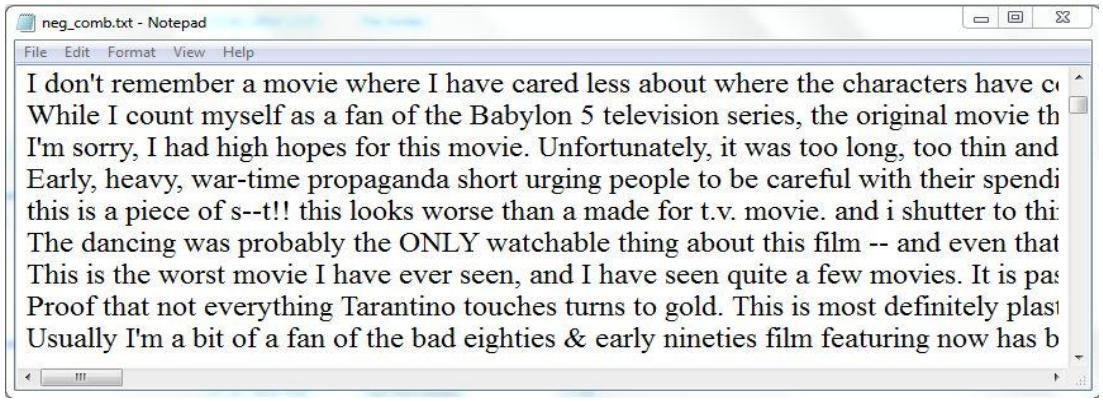

Figure 6.6: Positive Training Set

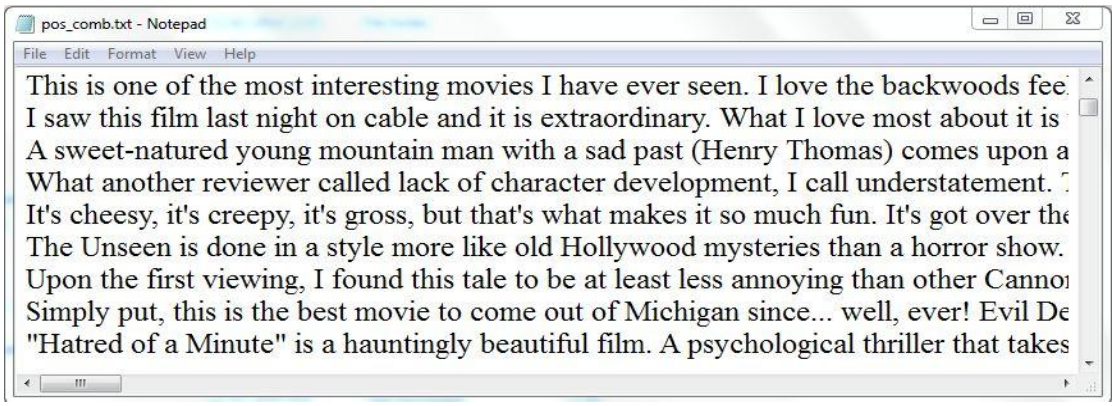

Figure 6.7: Negative Training Set

Using the predict function we specify the rows to be used as test data. Extracted data is manually labelled and stored in two separate files as negative and positive test data.

predicted = predict(classifier, mat[2300:2401,]);

Naive bayes algorithm computes the probability of the word being used as positive or negative. The test data has 101 number of tweets having 57 positive and 44 negative tweets, classified manually. After applying naïve bayes on collected data, the confusion matrix is obtained as shown in Figure 6.8. Here, first row represents that out of 57 positive tweets, the model predicts 42 to be positive and the rest i.e. 15 to be negative tweet. Similarly, in second row, out of 44 actual negative tweets 31 are predicted to be negative and remaining are predicted as positive tweets.

		PREDICTED	
		Positive	Negative
ACTUAL	Positive	42	15
	Negative	13	31

Figure 6.8: Confusion matrix after application of naive bayes on collected data

6.4 Application of Adjective Analysis

Once we are provided with the set of ambiguous tweets (false polarized set), adjective analysis on the data is applied. Intensity of the tweet is determined with the help of adjectives and adverb present in text. We have formed corpus of adjectives and adverbs with positive and negative polarity. Once the ambiguous tweet is selected for adjective analysis, the adjectives and adverbs in the tweets and their position is searched. Then adjectives and adverbs are matched with the collected corpus to find the nature of those adjectives and adverbs.

Few of the ambiguous tweets with adjective analysis are as follows:

- *#superman not bad at all to watch for one time*

As this tweet contains two negative words as ‘not’ and ‘bad’ so overall polarity will be positive and it will be added to positive set.

- *#supermanvsbatman quite boring as the first half doesn’t excites*

As this tweet contain one negative word in first part as “quite” and “boring”. Second part also has one negative and one positive as “doesn’t” and “excites”. So overall polarity will be negative and it will be added to negative set.

- *Doesn’t disappoint to marvel fan.....great one #supermanvsbatman*

This tweet will be classified as positive as it contains two negative words as “doesn’t” and “disappoint”.

Table 6.1 depicts the accuracy attained by the various model applied for sentiment analysis on IMDB movies reviews dataset. By combining the naive bayes and adjective analysis accuracy of 88.5% is achieved which can be improved further by adding more sentence forms. The corpus of adjectives can also be extended to add more number of words.

Table 6.1 : Accuracy of Various Models

Author and Literature	**Approach**	**Accuracy(%)**
Pang et al.[16]	Naive bayes Maximum entropy Support vector machine Support Vector Machine and regression metric labeling	81.5 81.0 82.9 66.3
Salvetti et al.[53]	Naive Bayes Markov model	79.5 80.5
Mullen and Collier[54]	Support vector machine	86.0
Beineke et al.[55]	Naive Bayes	65.9
Matsumoto et al. [56]	Support vector machine	88.3

Figure 6.9 represents the various maximum accuracy levels attained by different models. It is clear that naïve bayes and adjective analysis has the highest accuracy, 88.5%.

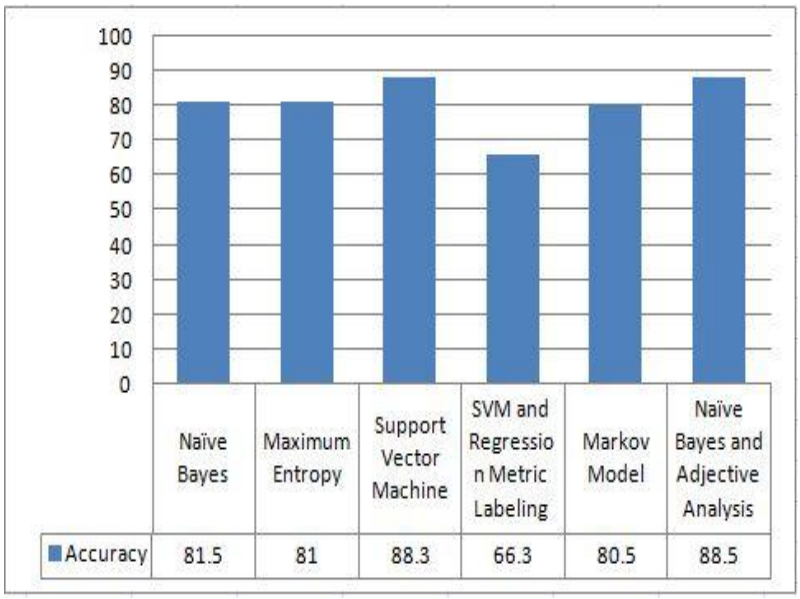

Figure 6.9: Graph representing Accuracy of various models

In this chapter we have evaluated the Integrated Naïve bayes and Adjective analysis and results are obtained step by step accordingly. We have different tools for its evaluation and testing. NodeXL is used for data collection, R and R Studio for pre processing and Naïve Bayes implementation and java for adjective analysis.

Chapter 7
Conclusion and Future Work

7.1 Conclusion

We have proposed an integrated model for sentiment analysis which combines Naïve Bayes and Adjective analysis. The model has four phases: 1. data collection; 2. preprocessing; 3. applying naïve bayes ; and 4. adjective analysis. The proposed model operates on collection of tweets related to movie reviews. Objectives of research have been fulfilled as follows:

We have studied various approaches for sentiment analysis like machine learning and natural language processing. We have studied various machine learning algorithms.

We have applied Naïve Bayes on collected tweets using R. The output of Naïve Bayes is further analyzed with adjective analysis.

We have compared the accuracy of our work with models available in research. The accuracy of our model has improved in comparison to the various combinations of models used by researchers. Results generate 88.5 % accuracy. Therefore, it can be deduced that sentiment analysis is enhanced by using combination of Naive bayes and adjective analysis.

7.2 Limitations

It works well with predefined types of sentences such as adjectives; verb plus adjectives; adverb plus adjectives. Another limitation is not considering the context in which the sentence is referred.

7.3 Future Scope

In future, the study can be enhanced by resolving some of the natural language problems like sarcasm, large number of adverbs and adjectives, and adding more forms of sentences. Corpus of adjectives and adverbs can be enhanced to achieve better accuracy.

Also the work could be carried out by employing NLP analysis to acquire wider approach and accurate output.

References

[1] H. Tang, S. Tan and X. Cheng, "A survey on sentiment detection of reviews", Expert Systems with Applications, vol. 36, no. 7, pp. 10760-10773, 2009.

[2] r. Leading social networks worldwide as of April 2016, "Leading global social networks 2016 | Statistic", *Statista*, 2016. [Online]. Available: http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. [Accessed: 26- May- 2016].

[3] T. Yanaru, T. Hirotja and N. Kimura, "An emotion-processing system based on fuzzy inference and its subjective observations", International Journal of Approximate Reasoning, vol. 10, no. 1, pp. 99-122, 1994.

[4] "How does sentiment analysis work? - Quora", *Quora.com*, 2016. [Online]. Available: http://www.quora.com/How-does-sentiment-analysis-work#!n=12. [Accessed: 26- May- 2016].

[5] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining.", LREc, vol. 10, pp. 1320-1326, 2010.

[6] 2016.[Online]. Available: http://www.stackoverflow.com/question/4806176/what-are-the-most-challenging-issues-in-sentiment-analysisopinion-mining. [Accessed: 27- May- 2016].

[7] B. Jansen, M. Zhang, K. Sobel and A. Chowdury, "Twitter power: Tweets as electronic word of mouth", J. Am. Soc. Inf. Sci., vol. 60, no. 11, pp. 2169-2188, 2009.

[8] M. Ghiassi, J. Skinner and D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network", Expert Systems with Applications, vol. 40, no. 16, pp. 6266-6282, 2013.

[9] Go, R. Bhayani and L. Huang, "Twitter sentiment classification using distant supervision", *CS224N Project Report, Stanford*, vol. 1, p. 12, 2009.

[10] Abbasi, H. Chen and A. Salem, "Sentiment analysis in multiple languages", *ACM Transactions on Information Systems*, vol. 26, no. 3, pp. 1-34, 2008.

[11] S. Das and M. Chen, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web", *Management Science*, vol. 53, no. 9, pp. 1375-1388, 2007.

[12] S. Tan, G. Wu, H. Tang and X. Cheng, "A novel scheme for domain-transfer problem in the context of sentiment analysis", *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 979--982, 2007.

[13] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences", *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 129--136, 2003.

[14] S. Ebert, N. Vu and H. Schutze, "Combining convolutional Neural Networks and SVMs for Sentiment Analysis in Twitter", SemEval-2015, p. 527, 2015.

[15] P. Turney and M. Littman, "Measuring praise and criticism", ACM Transactions on Information Systems, vol. 21, no. 4, pp. 315-346, 2003.

[16] Pang, L. Lee and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79--86, 2002.

[17] M. Hu and B. Liu, "Mining and summarizing customer reviews", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177, 2004.

[18] S. Kim and E. Hovy, "Determining the sentiment of opinions", Proceedings of the 20th international conference on Computational Linguistics, p. 1367, 2004.

[19] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff and S. Patwardhan, "OpinionFinder: A system for subjectivity analysis", Proceedings of hlt/emnlp on interactive demonstrations, pp. 34-35, 2005.

[20] M. Hu and B. Liu, "Mining and summarizing customer reviews", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177, 2004.

[21] S. Kim and E. Hovy, "Determining the sentiment of opinions", Proceedings of the 20th international conference on Computational Linguistics, p. 1367, 2004.

[22] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff and S. Patwardhan, "OpinionFinder: A system for subjectivity analysis", Proceedings of hlt/emnlp on interactive demonstrations, pp. 34-35, 2005.

[23] P. Turney and M. Littman, "Measuring praise and criticism", ACM Transactions on Information Systems, vol. 21, no. 4, pp. 315-346, 2003.

[24] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data", Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36--44, 2010.

[25] S. Hernandez and P. Sallis, "Sentiment-preserving reduction for social media analysis.",Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 409--416, 2011.

[26] R. Sharma, M. Gupta, P. Bhattacharyya and A. Agarwal, "Adjective Intensity and Sentiment Analysis".

[27] E. Kouloumpis, T. Wilson and J. Moore, "Twitter Sentiment Analysis:The Good the Bad and the OMG!", Icwsm, vol. 11, pp. 538--541, 2011.

[28] K. Liu, W. Li and M. Guo, "Emoticon Smoothed Language Models for Twitter Sentiment Analysis", AAAI, 2012.

[29] F. Benamara, C. Cesarano, A. Picariello, D. Recupero and V. Subrahmanian, "Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone", ICWSM, 2007.

[30] Harb, M. Planti{\'e}, G. Dray, M. Roche, F. Trousset and P. Poncelet, "Web opinion mining: How to extract opinions from blogs?", pp. 211--217, 2008.

[31] S. Mohammad, S. Kiritchenko and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets", arXiv preprint arXiv:1308.6242, 2013.

[32] W. Peng and D. Park, "Generate Adjective Sentiment Dictionary for Social Media Sentiment Analysis Using Constrained Nonnegative Matrix Factorization", *Urbana*, vol. 51, p. 61801, 2004.

[33] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach", Journal of Informetrics, vol. 3, no. 2, pp. 143--157, 2009.

[34] A.Barhan and A. Shakhomirov, "Methods for Sentiment Analysis of Twitter Messages",12th Conference of FRUCT Association, 2012.

[35] Y. Kim, "Convolutional Neural Networks for Sentence Classification", arXiv preprint arXiv:1408.5882, 2014.

[36] Pang and L. Lee, "Opinion Mining and Sentiment Analysis", *FNT in Information Retrieval*, vol. 2, no. 12, pp. 1-135, 2008.

[37] M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments", Expert Systems with Applications, vol. 40, no. 10, pp. 4241--4251, 2013.

[38] H. Kang, S. Yoo and D. Han, "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews", Expert Systems with Applications, vol. 39, no. 5, pp. 6000--6010, 2012.

[39] L. Chen, C. Liu and H. Chiu, "A neural network based approach for sentiment classification in the blogosphere", Journal of Informetrics, vol. 5, no. 2, pp. 313--322, 2011.

[40] Al-Daoud, "Integration of Support Vector Machine and Bayesian Neural Network for Data Mining and Classification", World Academy of Science, Engineering and Technology, vol. 64, p. 202, 2010.

[41] Kontopoulos, C. Berberidis, T. Dergiades and N. Bassiliades, "Ontology-based sentiment analysis of twitter posts", *Expert Systems with Applications*, vol. 40, no. 10, pp. 4065-4074, 2013.

[42] K. Yessenov and S. Misailovic, "Sentiment Analysis of Movie Review Comments", Methodology, pp. 1--17, 2009.

[43] J. Read, "Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification", Proceedings of the ACL student research workshop, pp. 43--48, 2005.

[44] S. Moghaddam and F. Popowich, "Opinion polarity identification through adjectives", arXiv preprint arXiv:1011.4623, 2010.

[45] O. Irsoy and C. Cardie, "Opinion Mining with Deep Recurrent Neural Networks", EMNLP, pp. 720--728, 2014.

[46] Montejo-Raez, E. Martnez-Cmara, M. Martn-Valdivia and L. Urea-Lpez, "Ranked WordNet graph for Sentiment Polarity Classification in Twitter", Computer Speech \& Language, vol. 28, no. 1, pp. 93--107, 2014.

[47] N. Li and D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast", Decision support systems, vol. 48, no. 2, pp. 354--368, 2016.

[48] Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs", arXiv preprint arXiv:1103.2903, 2011.

[49] J. Martineau and T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis", ICWSM, vol. 9, p. 106, 2009.

[50] K. Hiroshi, N. Tetsuya and W. Hideo, "Deeper sentiment analysis using machine translation technology", vol. 20, p. 494, 2004.

[51] https://nodexl.codeplex.com/. Accessed:2016-4-14

[52] “Text Classification Using Naive Bayes", *YouTube*, 2016. [Online]. Available: https://www.youtube.com/watch?v=EGKeC2S44Rs. [Accessed: 22- May- 2016].

[53] Salvetti, S. Lewis and C. Reichenbach, "Automatic opinion polarity classification of movie reviews", Colorado research in linguistics, vol. 17, p. 2, 2004.

[54] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources", EMNLP, vol. 4, pp. 412--418, 2004.

[55] P. Beineke, T. Hastie and S. Vaithyanathan, "The sentimental factor: Improving review classification via human-provided information", Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, p. 263, 2003.

[56] S. Matsumoto, H. Takamura and M. Okumura, "Sentiment classification using word sub-sequences and dependency sub-trees", Advances in Knowledge Discovery and Data Mining, pp. 301--311, 2005.

List of Publications

i. Mohit Mertiya and Ashima Singh, **"***Combining Naïve Bayes and Adjective Analysis for Sentiment Detection on Twitter" in* ***IEEE*** *International Conference on Inventive Computation Technologies (ICICT 2016) – reg.*

[accepted]

YouTube video link

The video link of the performed work can be accessed from

https://youtu.be/1pGp8GNWis8

Plagiarism Report

5%	3%	3%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Studies in Computational Intelligence, 2016. Publication	1%
2	balog.hu Internet Source	1%
3	www.ijretm.com Internet Source	1%
4	stackoverflow.com Internet Source	<1%
5	pureapps2.hw.ac.uk Internet Source	<1%
6	www.nit.eu Internet Source	<1%
7	He, Wu Guo, Lin Shen, Jiancheng Akula, V. "Social Media-Based Forecasting: A Case Study of Tweets and Stock Prices in the Financial Services In", Journal of Organizational and End User C, April 2016	<1%