SENTIMENT ANALYSIS OF TWITTER DATA USING MACHINE LEARNING TECHNIQUES

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

Master of Engineering

in

Software Engineering

Submitted By

Rohit Joshi

(Roll No. 801431022)

Under the supervision of:

Mr. Rajkumar Tekchandani

Assistant Professor

Thapar University, Patiala

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, "*Sentiment analysis of twitter data using machine learning techniques*", 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 *Mr. Rajkumar Tekchandani* 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.

(Rohit Joshi)

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

(Mr. Rajkumar Tekchandani)
Assistant Professor, CSED

Countersigned by

(Dr. Maninder Singh)
Head
Computer Science and Engineering Department
Thapar University
Patiala

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

Abstract

Online Microblogging on social networks have been used for indicating opinions about certain entity in very short messages. Existing some popular microblogs like twitter, facebook etc,in which twitter attains maximum amount of attention in the field of research areas related to product, movie reviews, stock exchange etc. The research on sentiment analysis has been going for a long time. Sentiment analysis in present days becomes the major issue in field of research and technology. Due to day by day increase in the number of users on the social networking websites, huge amount of data produces in the form of text, audio, video and images. There is need to do sentiment analysis as texts in form of messages or posts to find the whether the sentiment is negative, positive or neutral. We had extracted data from twitter i.e. movie reviews for sentiment prediction using machine-learning algorithms. We applied supervised machine-learning algorithms like support vector machines (SVM), maximum entropy and Naïve Bayes to classify data using unigram, bigram and hybrid i.e. unigram + bigram features. Result shows that SVM surpassed other classifiers with remarkable accuracy of 84% for movie reviews.

Acknowledgement

First of all I would like to thank the Almighty, who has always guided me to work on the right path of the life. It is a great privilege to express my gratitude and admiration towards my respected supervisor **Mr. Rajkumar Tekchandani** Assistant Professor Computer Science & Engineering Department. HE has been an esteemed guide and great support behind achieving this task. This work would not have been possible without the encouragement and able guidance of her. I also thank my supervisor for his time, patience, discussions and valuable comments. His enthusiasm and optimism made this experience both rewarding and enjoyable. I am truly grateful to him for extending his total co-operation and understanding whenever I needed help and guidance from him. I am also heartily thankful to **Dr. Deepak Garg**, Associate Professor and Head, Computer Science & Engineering Department and **Dr. Rupali Bhardwaj**, PG coordinator, for motivation and providing uncanny guidance and support throughout the preparation of the thesis report.

I will be failing in my duty if I do not express my gratitude to **Dr. S. S. Bhatia**, Senior Professor and Dean of Academic Affairs, for making provisions of infrastructure such as library facilities, computer labs equipped with net facilities, immensely useful for the learners to equip themselves with the latest in the field.

I am also thankful to the entire faculty and staff members of Computer Science and Engineering Department for their direct-indirect help, cooperation, love and affection, which made my stay at Thapar University memorable. Last but not least, I would like to thank my family for their wonderful love and encouragement, without their blessings none of this would have been possible.

Rohit Joshi

(801431022)

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Social Networking

Social networking is the grouping of individual into specific groups. It could be apolitical or religious group or group of college students, teenagers, all together sharing information of their interests, mostly online. Twitter, MySpace or Facebook are some of social networking sites that are free of charges and easy to access. This interaction is likely to include friendship, families, group relation and romantic ones. Social networking helps people to make new friends and develop some personal relationships and stay in touch with family very easily. Due to vast number of people connects to networking sites, number of relationships gradually increases. Social networking features combined in one website are: user groups, the latest info about music groups, places for videos and photos, blogs, personal profile, and much more. Social networking sites also helps people for maintaining and developing business contacts contact with them. LinkedIn is the best example for this, as it can be suitable place to talk about business and meet with professionals. It's easier and faster to be involving with new business clients. Internet is foremost and first communication technology with the capability to change social interaction of the people. Since early 1990s, adoption of internet has grown rapidly. For Example by 2003 63% of American had used the internet. In 1990s, Information technology experts expected the internet to be consigned to the trash heap of history.

Internet has become an essential part of our lives; many websites have facility ways for people to keep in touch in the form of social networking. Social networking sites are the way for interact with new people and to make connections as well as share photos, videos, and activities with each other

According to Amanda Lenhart and Mary Madden indicates that 55% of online teenagers have created a personal profile online, and 55% have used networking sites like MySpace and Face book. A social networking site includes both the exchange of information among individuals and group online. Expression also represents a view perspective, reflection, or quality of the individual or groups.

Figure 1.1 Abstract view of social networking

In 1997, first social site was launched named sixdegree.com. The intention of this site was to make online dating smoother, and the first time were able to allow users to create their personal profile and then post it online and even surf the network. After sixdegree.com, other social sites were launched, and served for a while, and failed to become a sustainable business entity. Since then Ryze.com (2002), match.com, were launched and used by many users. Uncontrolled use of microblogging separates the users from the real world life and creates shortage of attention. Use of social networking site has disadvantages and advantages. It is up to the interest and knowledge of the user to know how to use them, when and which site to use.

Social networking has some of the advantages like the meeting places but virtually where people can share thoughts with whoever they want and meet them in the first place. Some people used social networking sites to meet new friends, establish relationships and even marry, and have children. Some of them used this websites, in order to find their lost friends in their life and meet them. Use of social net working makes life faster and easier to get the latest news across the world at any moment and being updated. Nowadays,

Colleges and universities are getting fond of social networking, which make easier for faculties and students to find information freely and easily. Corporate and technical sectors also started recruiting employees with the help of social networking by go through with their profile and background. Since social networking sites are operating worldwide; it breaks some of the cultural barriers around the different part of the world.

Different people from different parts of the world can be able to connect with their loved one and families easily and without any cost. Social networking brings the world together and modifies communication. Everything in the world has both advantages and disadvantages even for social networking, the disadvantages are as follows; personal identity theft is the most popular one.

Social networking requires user's personal details in order to gain full access to the site as sign up. Recent information and news disclosed that some of the social networking websites misuses the personal information of users. Advertisers evade users' privacy. Sex offenders and criminals often visit the sites to find new victims. Some people mostly young ones for sake of revenge and hate post embarrassing information or photos, will have affect the future socially and mentally. These type of crimes called cyber bullying in social networking makes this much faster and easier, unfortunately sometimes even led to death of teens. The developers made the social networking sites for better communication but people rather addictive to those sites. This will hinder the ability of young people to develop real social life, face to face meeting of people, which is very important in developing conversation and speech. The traditional face to face socializing is becoming obsolete.

1.2 Online Microblogging

Online microblogging is broadcast medium that exists similar to blogging. Microblogging is different from blogging as its content normally smaller in both total and actual file size. Microblogs allow users to share small chunks of content such as video link, individual images or short messages, which may be the major reason for their popularity. These small messages are sometimes called microposts. As with traditional blogging, microbloggers post about topics varying from the simple theme such as "what

I’m doing now” to the particular theme like "most watched movie." Microblogs also used for commercial purposes to promote collaboration within websites, products and services and an organization.

Almost all the microblogging platforms offer features like privacy settings, in which users allow to control microblogs by selective ways of publishing entries along with the interface based on web or giving options of their chosen readers. These may include text messaging and instant messaging, E-mail, digital video and digital audio.

Microblogging is slowly moving into the mainstream. For Example, In The United States of America, Presidential candidate Barack Obama microblogged from the campaign trail using Twitter, one of the most popular microblogging services. Traditional organization of media, like The BBC and the New York Times, have begun to send links and headlines in microblog posts..

Advantages of Microblogging over traditional blogging:-

Why would anyone want to start posting on a microblogging site? If you've been hesitant to jump on on a site like Twitter or Tumblr, here are a few reasons to consider trying them

- **Developing content takes less time:** The traditional blogs are quite lengthy so that it takes time to complete our intent. Microblogging gives you the benefit of posting the most recently happened incident to aware their loved ones in shot time and message.
- **Individual parts of the content consumed in less time :** Hence microblogging is such a popular and interactive form of information consumption and social media on mobile devices, because as the gist of the content to the people increases , therefore it is best way where the news comes in short and precise way as compare to long ones that takes time.
- **Increases chances of frequent posts:** Microblogging involves the more frequent posts and shorter ones whereas traditional blogging involves exactly opposite less

frequent post and longer. Since you're saving so much time by focusing on just posting short pieces, you `can afford to post more frequently.

- **Share time sensitive or urgent information in a n easier way:** Huge number of the microblogging platforms have been made to be fast and easy to use. With aVine video, Tumblr post, Instagram photo or simple tweet, you can easily share to everyone on what's happening in your life or any news at this very moment.
- **Communication with followers becomes easy and direct :** In addition to communicate easily with greater short and frequent posts, microblogging platforms can be used to easily encourage and facilitate better interaction through liking, reblogging, tweeting , commenting and more.
- **Convenient using with mobile and tabs:** Microblogging gains too much of attention in present days and the main cause behind this is increasing trends of mobile browsing. It is difficult to consume, interact and write long and lengthy blog post in a tab or smartphone that's why microblogging comes into play and provide small, easy and faster posts .

1.3 Twitter

Twitter is an online microblogging service that allows users to read and write short sentences of length 140 characters called tweets. Twitter Inc. is located at San Francisco. Users should be register first to post any message, whereas unregistered users can only read them. Users can access Twitter with the website interface, mobile application or SMS. Twitter was created by Noah Glass, Biz Stone, Evan Williams and Jack Dorsey in March 2006 and launched in July 2006. Twitter has 310 Million monthly active user, 1 Billion Unique visits monthly to sites with embedded Tweets, 83% of active users are access through mobile application, consists of 3500 employees around the world, more than 35 offices across the world, 79% accounts are from outside U.S. , supports more than 40 languages and 40% employees of twitter are from technical background. All numbers approximate as of March 31, 2016.

The company experienced rapid initial growth. In year 2007 around 4,00,000 tweets was posted per quarter. In 2008 this extends to 10 million tweets a quarter. 50 millions tweets

were posted per day in February 2010. 70000 application were registered by company as March 2010.

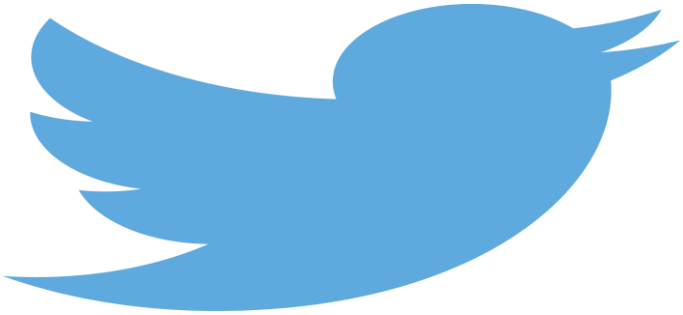


Figure 1.2 Twitter Logo

According to Twitter 750 tweets posted each second which equals to each day around 65 million tweets were posted as of June 2010. On daily basis around 140 tweets were posted by March 2011. In January 2009, since it gained lot of popularity, Twitter becomes third-highest online microblogging site, given by Compete.com.

The reason we are using microblogging and twitter data are following:

- The scope of microblogging tends to grow bigger and bigger day by day. Easy to use and people can share and give opinions on certain topic, thus it makes essential source.
- Twitter generates vast number of messages that is increasing exponentially. The extracted data can be enormously large.
- Twitter users varies from person to person as user can be politician, film stars, celebrities, sportsman and many leaders across the country including prime minister of India. So it contains all the messages of different caste, religion and sex.
- Twitter users are all over the world so it contains data for different language.

1.4 Sentiment Analysis

Sentiment Analysis is to determine the opinion of user related to some event or the statement describe the emotion of the user i.e. what he/she feel about it. Users share the

things about their ongoing life, discuss current issues and variety of topics. Independent to write in any format without following rules that makes this more popular than older blogging sites. Movies and product reviews easily available now a days or thoughts on religious and political issues, so it becomes essential sources of user sentiment and opinion.Data that we using in our experiment are from twitter, it contains vast number of messeges by large number of users created by themselves. Messeges can vary from public opinion to personal thought. As an example some post from twitter can be shown in Table 1.1.

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. Users are increasing day by day as the population and trend of using microblogging sites are increasing, so the data can be used in research purpose of sentiment analysis and opinion mining.

For example, movie makers interested in following questions:-

- What is audience expectation from our movie?(whether the movie is likable or not)
- How the people reacted to our movie?
- Whether the movie is turn to be good or bad?

In the time of election every news channel show the exit polls of every political parties, so every political party willing to know how many are in favor and with the help of microblogging sites people will give the opinions about likes and dislikes of the party. These opinions will help parties to increase their voters.

The data we using in this experiment are movie reviews. We have collected about 17000 movie reviews from twitter. The movie reviews contains reviews of different movies. Reviews can be categorized in three ways:

1. Positive reviews: messages in which people liked the movie.
2. Negative reviews: messages in which people not liked the movie.

3. Neutral reviews: messages in which people doesn't have any emotion or based on mere fact.

We have extracted 5000 each positive, negative and neutral reviews for training set and 2000 reviews will be used in test set. We show comparison between the different machine learning classifiers and find out which will give best results among these.

Table 1.1: Examples of positive, negative and neutral tweets

Sentiment	**Tweet**
Positive	The creators of south park in their own film here, this is a brilliant film with a huge entertainment factor. If you like Naked Gun films and are not young and not too mature or serious on your humor, you'll love this.
Positive	This is the definite Lars von Trier Movie, my favorite, I rank it higher than "Breaking the waves" I simply love the beauty of the picture...the framing is so original; acting is wonderful, A MUST SEE.
Negative	Long, boring, blasphemous. Never have I been so glad to see ending credits roll.
Negative	I rated this a 3. The dubbing was as bad as I have seen. The plot - yuck. I'm not sure which ruined the movie more. Jet Li is definitely a great martial artist, but I'll stick to Jackie Chan movies until somebody tells me Jet's English is up to par.
Neutral	Love to watch commercial movies only in free time but get bored easily.

Chapter 2: Literature Survey

Sentiment Analysis is the thorough research of how opinions and perspectives can be relate to ones emotion and attitude shows in natural language respect to an event. Recent events show that the sentiment analysis has reached upto great achievement which can surpass the positive vs negative and deal with whole arena of behavior and emotions for different communities and topics. In the field of sentiment analysis using different techniques good amount of research has been carried out for prediction of social opinions.

Pang and lee [1] proposed the system where an opinion can be positive or negative was found out by ratio of positive words to total words. Later in 2008 the authors developed methodology in which tweet outcome can be decided by term in the tweet. Compare to baselines that are generated by humans, the results are pretty good when machine learning techniques are used. SVM gave best result as compare to Naïve Bayes. Regardless of using different types of features the authors did not attain desired accuracies over topic based categorization.

Jiang *et al.* [2] focus on target-dependent sentiment classification. Here target-dependent means whether the sentiment is positive, negative or neutral depends on nature of the question that is asked. The authors proposed to make better target-dependent sentiment classification by joining features of target-dependent and considering related tweets. The authors also proposed that there is need of consideration current tweets to the related tweets by employing graph based optimization. As claimed by authors experimental results, the graph based optimization increases the performance.

Tan *et al.* [3] said that the users that shared similar opinions are likely to be connected. The authors proposed the model that were generated from either by following the network that has been made by tagging different user with the help of "@" or by analyzing the network of twitter follower/followee. The authors explained that by employing information of link of twitter there will be improvement in user-level sentiment analysis.

Chen *et al.* [4] employed the feed-forward BPN network and uses sentiment orientation to calculate the results at each neuron. The authors proposed a methodology based on neural network. The proposed methodology is combination of machine learning classifiers and semantic orientation indexes. In order to obtain efficiency in methodology, semantic orientation indexes used as inputs for neural network. The proposed methodology outperforms other neural networks and traditional approaches by increasing efficiency in both training as well as classification time.

Malhar and Ram [5] employed supervised machine learning techniques and artificial neural networks to classify twitter data along with case study of Presidential and Assembly elections which results SVM outperforms all other classifiers. The authors proposed a methodology to predict the outcome of election results by utilizing the user influence factor. To carry out reduction in dimension the authors combined the Principle Component Analysis with SVM.

Anton and Andrey [6] reviewed the existing techniques and developed a model for automatic sentiment analysis of twitter messages using unigram, bigram and jointly i.e. hybrid feature. The purpose of the authors is to explore and produce approaches for analyzing the accent of the messages in social media. The authors reviewed existing automatic sentiment analysis approaches and in order to maintain the context of growing methods the character feature of social media statements were studied.

Pak and Paroubek [7] perform linguistic analysis and build a sentiment classifier to determine positive, negative and neutral sentiments for a document. The authors developed a sentiment classifier, that gives neutral, negative and positive statements of a document. In order to train sentiment classifier the author proposed an approach that collects corpus automatically. In order to analyze the dissimilarity in diffusion among neutral, negative and positive sets, the authors used TreeTagger.

Kopel and Schler [8] explain that it is very important to use neutral messages to get good knowledge of polarity. The authors also states that positive and negative messages alone will not give proper understanding about neutral messages. Knowing about neutral messages clear the difference between positive and negative messages. The authors found

that in one of the corpus having most of the neutral documents gives no sentiment which can be used as counter for testing both positivity and negativity of a document.

Go *et al*. [9] introduced a methodology for automatic sentiment classification of twitter messages. Respective of query term messages were classified as negative or positive. Here authors uses distant supervision to display the results of sentiments of twitter posts with the help of the machine learning algorithms. The algorithms such as Maximum Entropy, SVM and Naïve Bayes are applied to training data which contains emoticons, gave accuracy above 80%. The authors also discuss about preprocessing steps that was helped to obtain higher accuracy. The authors came up with an idea for distant supervised learning using tweets that contain emoticons.

Christianini and Taylor [10] published and shared the knowledge about SVM which is machine learning algorithm. The authors manage to give deep understanding about algorithm and how to approach the SVM algorithm in order to implement it to solve the practical problems. The approach will be theoretical as when the book was published, the research was on going on every field.

Burger *et al.* [11] Since, In this era the computer have become enough powerful that can handle large scale application which gives pattern recognition and statistical estimation of real world problems. The authors introduced a n approach for statistical modeling based on maximum entropy. By using examples of problems in natural language processing, the authors shows maximum-likelihood methodology for automatic construction of maximum entropy models. Here the authors described the principle of maximum entropy. This principle selects the model with greatest entropy among all the consistent models. By maximizing the likelihood of training data we can obtain optimal values of given parameters.

Romero *et al.* [12] discovered that hashtags becomes the common feature of twitter used in every message and new terms are created and changing on daily basis which effects the general meaning of the original term. The authors also found structural difference among issues and learn the structure of widely used different types of hashtags. The

authors also developed generative and simulation based models to study the interaction between design of latest adopters on which hashtag expands and adoption dynamics.

Li and wu [13] stated emotional polarity computation as sentiment analysis which have become prospering boundary in the community of text mining. With the help of text mining and sentiment analysis, here the authors studied about hotspot detection and forecast. The authors created an algorithm which describes emotional polarity of a message and obtain a value of each word in it. To create unsupervised text mining method, this algorithm is combined with support vector machine (SVM) and K-means clustering. After the experimental study both K-means and SVM obtain the same results for top 4 hotspots of the year.

Tan and Zhang [14] Until this date very less number of researches carried for the Chinese documents on sentiment analysis. The authors studied sentiment categorization on Chinese documents. The selection methods were featured as Document Frequency (DF) , CHI, Information Gain (IG) and Mutual Information (MI). The machine learning methods that are used are support vector machine (SVM), Naïve Bayes (NB), Winnow classifier, K-nearest neighbor classifier and Centroid classifier. Size of 1021 Chinese document were investigated. For selection of sentiment terms Information gain (IG) performs best and for sentiment classification SVM outperforms all the classifiers.

Martineau and Finin [15] proposed a technique called Delta TFIDF which measure word scores efficiently before classification. Delta TFIDF was easy to understand, implement and compute. For sentiment classification the authors used support vector machines to achieve better accuracy with Delta TFIDF and using data sets of movie reviews. The authors said that Delta TFIDF is better than TDFIF feature and count term raw for all sizes of documents that weights for congressional detecting support for bill, sentiment polarity classification and subjectivity detection. The authors stated Delta TFIDF is first measuring approach to boost and identify the relevance of selective words using the calculated unsupervised distribution of features before classification between the two classes.

Nielson [16] developed a labeled word list in which scores of the effective words had been obtained comes into the messages analyzing for sentiment analysis. Before the arrival of sentiment analysis and micro blogging there contains an effective term list for e.g Affective Norm for English Words (ANEW) developed by the author. The author made the word list exclusively for micro blogs i.e. ANEW in comparison with other list which can be used for detecting sentiment strength for micro blogs. The author used Twitter posted messages for scoring words for sentiment analysis.

Mohammad *et al.* [17] developed two SVM classifiers, one is term level task which determines the sentiment of a word in the message and one is message level task which determines the sentiment of messages such as SMS and tweets. The authors took part in a competition where 44 teams came in their submissions stood fist in work on tweets, getting 88.93 F-core in term-level task and 69.02 F-score in message-level task. The authors executed sentiment, semantic and surface-form features. The authors also produced two big term-sentiment associations, first with emoticons from tweets and second with sentiment –term hashtags from tweets.

Kouloumpis *et al.* [18] explored the advantage of semantic features for analyzing the sentiment of messages of Twitter. The authors investigate the features that collects knowledge about intuitive and informal language that used in microblogging as well as advantage of existing lexical resources. The authors used the supervised learning method to the problem and to collect it hashtags are used. The authors concluded that in the experimental study part-of-speech feature not better for sentiment analysis when it comes to the domain of miocroblogging on twitter and it confirmed that for collecting data hashtags are very useful so that messages with negative and positive emoticon.

Denecke [19] proposed an approach for deciding polarity of word in framework of multilingual. The approach influences on lexical resources available in English for sentiment analysis. In this approach first the language itself is translated into English using the standard translation software. Further in translation the document is then classified into positive and negative class for sentiment analysis. The classification can be done on the basis of the adjective present in the document. The authors concluded that it is feasible approach to sentiment analysis in the multilingual framework.

Gokulkrishnan *et al.* [20] proposed a methodology for the preprocessing of publically generated tweets from twitter online microblogging site and on the basis of their opinion content of irrelevant, negative or positive sentiment classified can be done; and investigating the performance different classifying methods based on precision and recall. The authors explained limitations and applications of the research. The authors also handled the skewness of the datasets by exclusively new approach called SMOTE oversampling method which helped by increases the accuracy of the classifier. Random Forest, SVM and SMO generates better performance compare to Naïve Bayesian classifier.

Neri *et al.* [21] performed sentiment analysis on newscast over more than 1000 Facebook posts and then compared the sentiment for dynamic company La7 and Rai – the Italian social broadcasting company which is emerging company. The authors observations were mapped with the study conducted by the Italian research institute highly specialized in study of media at empirical and theoretical level, occupied in the study of communication of politics in the mass media known as Osservatorio di Pavia. The authors experiment done by Knowledge Mining System which is used by security related agencies and institution of government in Italy to control information contained Web Mining and OSINT.

Wilson *et al.* [22] said that the methodologies for automatic sentiment analysis start with a big set of terms noted with their respective polarity. The main purpose of this study is to easily differentiate between contextual and prior polarity, with prior knowledge of understanding which are the necessary features for this task. The experiment covers the feature performance for multiple algorithms of machine learning. Except one algorithm, features when combined together gives the best results. The evaluations shows that when natural instances are present the performance of features degraded on great pace. The authors suggested that indicating features that described more complex interdependencies between polarity clues can be considered as future research work.

Godbole *et al.* [23] proposed a system which contains phase of identification sentiment in which for a particular topic which displays some opinions and scoring phase and a sentiment aggregation that will scores relative entities in the same class. At last the

authors investigates the importance of scoring methods over big dataset of blogs and news. The authors interested in the fact that sentiments can vary according to the geographic location, news source or demographic group. As future work the authors are studying in evaluating the extent to which we predict the changes of future in behavior of market or popularity.

Benamara *et al.* [24] stated that most of the work done in past is finding the strength of subjective statements within a document or expressions uses the special part of speech such as nouns, verbs and adjectives. The authors said that until their contribution there was not a single related to adverbs in sentiment analysis nor use of adverb-adjective combinations (AACs) in sentiment analysis. The authors proposed a sentiment analysis method which is based on AACs which uses a linguistic evaluation of degrees of adverbs. The authors explained the experimental results on dataset of 200 news articles and compares the proposed technique with existing techniques of sentiment analysis. Based on Pearson correlation with human objects their experimental results gives higher accuracy.

Boyd and Ellison [25] stated that social networking sites (SNSs) are regularly seeking the attention of industry and academic researchers fascinated by their reach and affordance. The authors described in the introductory article the functions of SNSs and introduce a complete definition. The authors presented an aspect on the history of such sites, explaining development and key changes. The authors finally concluded that the condition is changing drastically and people should aware of which sites is using and why and for what purposes, especially other countries than U.S.

Agarwal *et al.* [26] performed sentiment analysis on twitter data. The authors proposed functions polarity prior POS- specific and studied the usage of a tree kernel to prevent the necessity for hectic feature engineering. The tree kernel and the new functions performed approx at the same level both surpassing the state of the art baseline. The authors concluded that for twitter data sentiment analysis is not that different as sentiment analysis for different genres.

Nasukawa and Yi [27] proposed an approach for sentiment analysis to find sentiments connected with negative or positive polarities from a document for specific subject, rather than classifying the whole document into negative or positive. The major problems in sentiment analysis are whether the statements points negative or positive behavior towards the subject and to be found how sentiment are described in texts. The authors stated that it is essential to clearly find out the semantic relationships between the subject and the sentiment expressions to increase the accuracy for the analysis of sentiment. In order to identify the sentiments in news articles and web pages, their proposed system obtained high precision of 75-95%.

Wang *et al.* [28] proposed a system in U.S. elections 2012 for presidential candidates using real-time evaluation of sentiment on online microblogging site twitter. In order to collect the poll data the traditional analysis of election takes much time, but with the help of this system it takes data from more people with help of twitter, a microblogging service. It helps the social people like scholars, media and politician to broadcast their future perspective of the public opinion and electoral process. The authors finally concluded that the system and approach are generic, and should be adopted easily and spread across various other domains.

Wilson *et al.* [29] presented a method which first describes the whether a statement is polar or neutral to phase-level sentiment evaluation and then ascertain the polarity of polar statements. By applying this methodology, the system is capable to identify automatically the contextual polarity of sentiment statements for huge subsets, obtaining results which are greater than baseline.

Kanayama and Nasukawa [30] proposed an unsupervised lexicon building approach which detected the clauses of polar that grant negative or positive effect in a particular domain. The entries that are lexical in nature to be received are called polar atoms, the lesser human-recognizable semantic models that justify clause polarity. By the usage of precision and overall density of consistency in the dataset, the statistical approximation selects necessary polar atoms through candidates , without change in the threshold values. The obtained result shows that the applied method is robust enough for datasets

with different domains and also for weight of initial lexicon and the precision of polarity report from the automatically received lexicon was on average of 94%.

Choi and Cardie [31] studied that the essential cooperation in event of compositional semantics and presents a learning based technique that connects structural aasumption by compositional semantics for learning method. The authors conducted experiments that shows compositional semantics based on natural heuristics that can outperform the learning based techniques which does not integrate compositional semantics, whereas a technique which consolidate semantics compositional onto learning which is greater that other all alternatives. The authors also studied that for describing expression-level polarity, content word negator plays an important role. Finally the authors concluded that accuracy of classification of expression level linearly decreases as context that is gradually determined.

Melville *et al.* [32] presented a uniformed framework with respect to world-class associations using background lexical information and improve the information by using one of the available training examples to a particular domain. Experimental results shows that the authors methodology better performs than using training data or background knowledge within separation and text classification with lexical knowledge using to optional methodology. The authors concluded that they made two contributions. Firstly, they described a uniformed framework for combining knowledge of lexical for categorization of text in supervised learning and secondly, successfully applied the described methodology to analysis of classification of sentiment.

Paltoglou and Thelwall [33] stated that a large number of sentiment analysis methodologies have used support vector machines as their baselines with the weights of binary unigram. The authors in this paper explored if there is any reliable feature weighted schemes which can improve accuracy of classification with the help of retrieving the information. The authors shows that alternatives of the *tf.idf* scheme gives notable increase in accuracy for sentiment analysis, with the use of sublinear function for smoothing of document frequency and term frequency weights. The methodology was tested on large data set and obtained highest accuracy.

Fernandez *et al.* [34] developed a system introduced for the Subtask B Sentiment analysis of twitter i.e. SemEval 2014 task9. The authors system comprises of supervised methodology using techniques of machine learning, which using the text in dataset as features. This work is totally independent of any external resources and knowledge. The originality of author methodology depends on the use of skipgrams, n-grams and words as features. In the experimental study, it is clearly proves that skipgram shows better results than the ngram or word for the given datasets.

Mullen and Malouf [35] developed initial tests of statistics on a fresh datasets postings of group of political discussion that indicates the post that made response direct to post of others that having a greater likelihood that presents the perspective of opposing politics that of original post. The authors concluded that's the approaches of traditional text classifications is insufficient for this task in this dataset of sentiment analysis and the improvement can be made by utilizing information about how posters cooperate with one another.

Harb *et al.* [36] stated that the previous approaches until this paper were written suffered from drawback i.e. for a particular topic either the adjective is not available or from another topic it meaning is different. The authors proposed a new methodology which consists of two steps. Firstly, for a particular topic the authors extract a learning dataset from the internet. Secondly the authors extracting from the dataset, they made two classes that are negative and positive adjectives with respect to the topic. The experimental study on the real dataset shows the importance of authors methodology. The experiments are performed on dataset that are cinema reviews and blogs shows that with the author methodology, it is easy to extract the desired adjectives for a particular topic.

Kim and Hovy [37] stated that the identification of a sentiment was challenging problem. The authors developed a system for a particular topic it automatically search the users who posts their views on that topic and the sentiment of each views. The systems consists a module for describing sentiment of a word and other for merging the sentiments into a statement. The authors did experiment with different models of classification and merging the sentiment at sentence and word level, given better results. For the improvement of recognition of Holder, the authors are using parser to attach areas that

are more reliable with Holders. The learning techniques that are used is this system are support vector machines and decision list.

Martalo *et al.* [38] investigated how factors that are affective impact on the dialogue patterns and whether this impact may be explained and identified by Hidden Markov Models (HMMs). The goal of the authors is to study the chance of applying this model to classify behavior of users for the purposes of adaptation. The authors obtained the initial results of their research and present a debate of problems that are open. With the help of the results, the author claims that the complicated interaction between the pragmatic level and the acoustic level comprises an important facet of emotions contained in voice expressions.

Daoud [39] proposed a classifier and the introduced classifier contains four components which are AdaBoost which is a piece of an algorithm, Bayesian neural network, support vector machine and a technique for feature selection that is Signal-to-Noise. To confirm the efficiency of introduced classifier, the authors applied seven traditional classifiers to four datasets. The experimental study shows that applying the introduced classifier increases the rates of classification for all datasets. The author stated that SVMs key features are the control over capacity attained by margin optimization, sparseness of solution, the lack of local minima and the usage of kernels.

Yessenov and Misailovic [40] presents study of effectiveness of techniques of machine learning in text message classification by semantic meaning. The authors use comments of movie reviews from Digg that is social network which is popular as authors dataset and text classification can be done by negative or positive and objectivity or subjectivity attitude. The authors suggested different methodologies in text feature extraction such as using knowledge of WordNet synonyms, bounding word frequencies by threshold, handling negations, restricting to adjectives and adverbs, using large movie review corpus and a bag-of-words model. The authors analyze their performance on accuracy using four methodologies of machine learning that are K-Means clustering, Maximum Entropy, Decision Trees and Naïve Bayes. Finally, the authors concluded that bag-of-words model perform better than relative models.

Kang *et al.* [41] stated that the senti-lexicon existed does not properly adopt the word sentiment used in the restaurant review. The author introduced a senti-lexicon of restaurant reviews for the sentiment analysis. Using supervised learning technique a review document is classified as negative sentiment and positive sentiment, hence there is chance for the accuracy of positive classification to greater than 10% than the accuracy of negative classification. The author also introduced an improved version of Naïve Bayes to deal with these types of problems. The authors improved Naïve Bayes had managed to low the gap between positive and negative accuracy by 3.8% when applied with unigram + bigram and 28.5% when compared with SVM.

Chapter 3: Research Problem

3.1 Problem Statement

Microblogging is type of blogging which consists of limited number of words. Limitation of words determined by respective microblogging sites. It gives right to share his/her thoughts, opinions and sentiments in less number of words. It is one of the revolutionary thing happened in the world of technology. People in these days depends upon microblogging sites such as twitter, facebook, tumblr etc. to communicate with both relatives and rest of world. Here sentiments comes into the play which will be shared by anyone in the time they feel and wanted to be shared. Sentiments are nothing but feelings respect to event. Sentiment Analysis is to determine the opinion of user related to some event or the statement describe the emotion of the user i.e. what he/she feel about it.

The research on sentiment analysis has been going for a long time. Sentiment analysis in present days becomes the major issue in field of research and technology. Due to day by day increase in the number of users on the social networking websites, huge amount of data produces in the form of text, audio, video and images. There is need to do sentiment analysis as texts in form of messages or posts to find the whether the sentiment is negative, positive or neutral.

3.2 Gap Analysis

A lot of research has been done in the area of sentiment analysis. Many researchers used Part-Of-Speech and polarity based feature using supervised learning techniques for classifying. Many automatic classifiers are proposed for classifying the texts in the given expressions but with the restricted domains, but there will be new informal words that are added to the present world which means something in the common social network, so there is need to include all the common referred terms that are used in the social networking world.

3.3 Objectives

The objectives of the thesis has been discussed in the following points :-

1. To explore, analyze and study the existing sentiment analysis detection techniques in the online microblogging network.
2. To study how the tweets can be generated from the twitter with the help of Java API.
3. To implement and analyze the results achieved after applying the supervised
4. Learning classifiers to the data set.

3.4 Research Methodology

The main aim of the thesis is to compare the results that are implemented with the help of supervised classifier

The methodology followed is:

1. We have collected a corpus of positive, negative and neutral tweets with the help of Twitter4j java API from Twitter. The size of our corpus can be enormously large.
2. We then remove the stop words from the collected corpus to make the content free from commas, full stops etc.
3. We then apply machine learning algorithms to our training set first and then test set and compare the results.

With the help of results we evaluate which machine learning algorithm is best for classification of sentiment Analysis.

Chapter 4: Methodology

4.1 Preprocessing

4.1.1 Collection of data

We collected data from Twitter API named as Twitter4j using netbeans. Searched given by using #Hashtag followed by the movie name like #FAN, #Bajarangi Bhaijaan, #The Jungle Book etc. Approx 17000 tweets have been collected from the various movie tweets.

Reviews can also be searched by #Hash tags followed by respective movie stars, directors, production house and music record companies. In twitter hash tags becomes the necessary symbol to find about something and it gives user limit of 140 words to express their views and attitude.

4.1.2 Normalization

We have found that to get desired results from the classifier we have to make sure that the tweets can be processed properly. As tweets can be in user language, so we have to clean every data which are irrelevant to the data. The following things which can be irrelevant to the data are:-

- URL's: URL's in the message will not make any sense as it simply distracts the result of classifier.
- Username: Removal of username can be necessary for cleaning purposes as it can effect falsely to our results.
- Repeated characters: If the character is repeated more than two time then it can be comprise new word but the meaning is same, so we have to eliminate that word and make the word genuine. For example good can be written as goooooood.

Repeated words : If the message contain word which has been appeared more than two times continuously then it has to be change into two times. For example great great great great movie can be covert to great great movie.

4.1.3 Removal of stop words

Stop words are the words like “a”, ”is”, ”the”, “etc” etc; These words has nothing to do with the emotion , so has to be discarded from the message. Now next step is to train the data using supervised classifier.

4.2 Machine Learning Techniques

We employed classification methods which is polarity based using set of positive, negative and neutral tweets provided by Twitter4j API. Polarity is given by ratio of probability of a word appeared in set of positive or negative statement which makes the word positive or negative. The classifiers we are using are based on the concept of polarity.

$$\text{Polarity} = \frac{P(\text{Postive\_Words})/P(\text{Total\_Words})}{P(\text{Negative\_Words})/P(\text{Total\_Words})} \quad (1)$$

If the feature is independent and based only on Standard English Dictionary then only this technique works. This method fails when we tried to record the sentiment shown with respect to comparison. Further, the polarity based technique also fails to record query related sentiment. In order to fulfill the requirement of classification we involved machine learning techniques.

The machine techniques comprised of following supervised classifier that are given below:-

- Naïve Bayes
- Support Vector Machines (SVM)
- Maximum Entropy (MaxEnt)

4.3. Supervised Classifiers

4.3.1. Naïve Bayes*:* The Naïve Bayes classifier in one of the simplest probabilistic model works positively on text categorization and employed on Bayes rule with self-supporting feature collection [3] works positively on text categorization and employed on Bayes rule

with self-supporting feature collection [3]. It is flexible in way of handling with any number of classes or attributes. For a given tweet d, C* is a class variable which defines the sentiment given by

$$C^* = argMax_c P_{NB}(C|D) \tag{2}$$

Bayes Probabilty $P_{NB}(C|D)$ described as

$$P_{NB}(C|D) = \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d)})}{P(d)} \tag{3}$$

Here, f is feature and $n_i(d)$ is feature count found in d, m represents total number of features and P(c) and P(f|c) are found through maximum likelihood estimates[8].

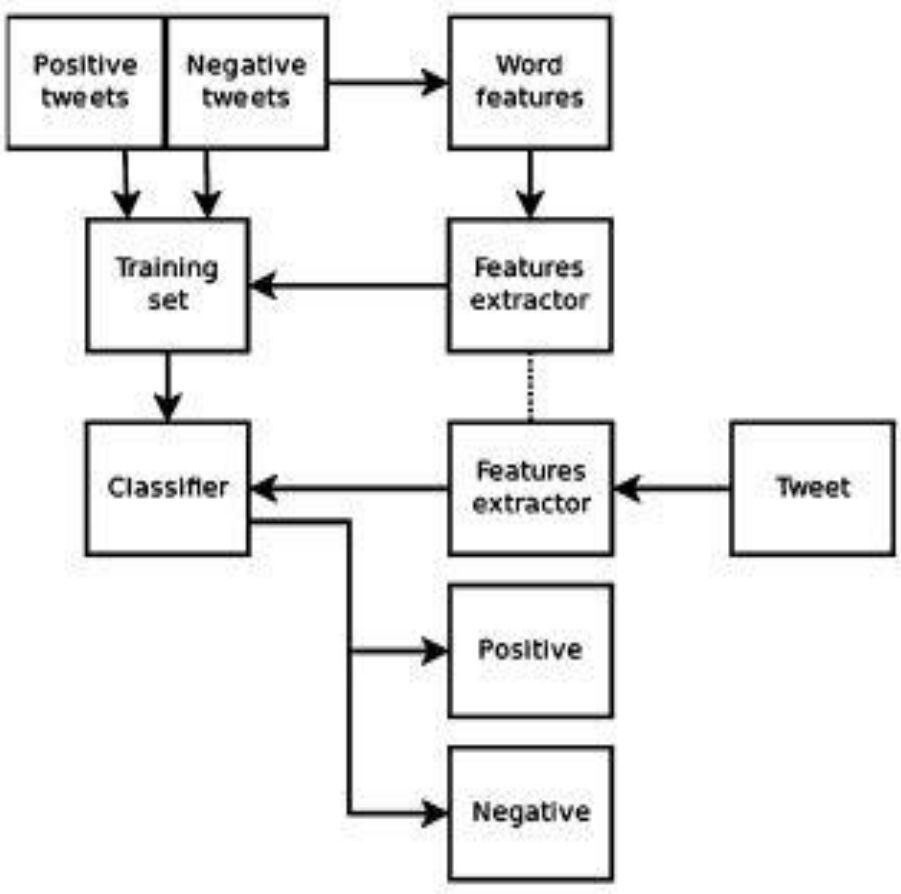


Figure 4.1: Flow Diagram of Supervised Classifiers

During classification phase we found a word which was not found in training phase then we will give zero as probability for positive, negative and neutral classes. To end this problem we tend to make probability equal using Laplacian smoothing constant k=1.

$$\frac{term\_count+k}{Total\_Terms+k|c|} \tag{4}$$

4.3.2. Support vector machines*:* SVMs are happen to be extremely accomplished at text categorization, widely outperforming Naïve Bayes (Joachims 1998). We examined big margin classifier to attain effective accuracy of classification process [9].

$$\vec{w}=\sum_i x_i c_i d_i \, , x_i \geq 0 \tag{5}$$

SVMs uses a function called kernel which are machine learning classification methodology in which the data is not separable linearly in the new area which it is to locate to area of data points, with allocation for classification of erroneous.

Support Vector Machines are the members of the family of classifiers which are linear. The main objective of the linear classifier is to find a hyperplane which is linear in nature of a feature area that divides all other entities in form of two classes. The main function of the SVMs is find out the hyporplane which is separating that has distance maximum from the nearest points to feature area in it.

Searching hyperplane in sample of linear separable, the equation can be consider as problem of optimization:

$$\frac{1}{2}\|\omega\|^2 \rightarrow \min(\omega, b \tag{6}$$

$$y_i(w^T x_i + \ b) \geq 1, j = 1, \dots, m, \tag{7}$$

Here, $\frac{1}{\|\omega\|}$, is area between the points of both second and first class and the hyperplane and it is nearest to $y_i(w^T x_i + \ b)$, the product of its position relative to the hyperplane and point class value.

The kernel we are using is linear, the parameters are all set to its default values, inputs of SVM are data vectors that has to fed in sets.

4.3.3. Maximum Entropy*:* MaxEnt is another classification technique widely used lot in applications of natural language processing [10]. MaxEnt not always but sometimes outperforms the Naïve Bayes classifier for text classification [11]. MaxEnt is the most uniform model prefer for the classification purpose [12].

In the scenario of 2-class, to search for distribution over the both classes it is likely the same thing as using the logistic regression. Regarding independence of feature, it does not make any assumption. Due to this we can add features and phrases such as bigrams and to MaxEnt without affecting overlapping of the features. Let's take an example in which we have two features such as "good" and other one is "very good" , then in case of Naïve Bayes their probabilities will be taken as independent even when the both of this are overlapping but not in case of MaxEnt. The equation for this model can be given as:

$$P(C|D) = {}^{1}/_{Z(d)} \exp(\sum_i \lambda_{i,c} F_{i,c}(d,c)) \quad (8)$$

Here, c indicates class, d indicates a single tweet, λ indicates vector of weight, normalization function Z(d), $F_{i,c\varphi}$ is a class/feature function for class c and feature f_i defined as follows :

$$f_{i,c}(d,c') = \begin{cases} 1, & n_i(d) > 0 \ and \ c' = c \\ 0, & otherwise \end{cases} \quad (9)$$

The vectors of weight decided the importance of feature classification. If weight is higher than it means the feature is the strong indicator of that class and vice versa.

4.4 Performance Measure

To calculate the accuracy of classifier we required measure on which accuracy can be obtained. There are two measures on which accuracy can be dependent:

- Precison
- Recall
- Accuracy

Let's take collection of M documents, M_P denotes the number of document which belongs to the true positive class and M_N denotes the number of documents which belongs to the true negative class. TP documents had rightly classified whereas FP documents are wrongly classified, similarly FN documents are wrongly classified and TN documents are rightly classified.

Precision: It is the ratio of documents of rightly classified under positive prediction class to all documents under positive prediction class.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Recall: It is the ratio of documents of rightly classified under positive prediction class to the documents that are positive in the negative prediction class.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

		True Class	
		Positive	Negative
Prediction Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Figure 4.2: Confusion Matrix

Accuracy: In order to check which n-gram feature will give better results for these three models, we have to find the accuracy of classifiers. Accuracy for any prediction model can be given as:-

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

.

Chapter 5: Implementation and Results

5.1 Implementation

5.1.1 Data Extracting

We are extracting tweets from the twitter with the help of the Java API called Twitter4j. It consists various number of libraries that are used in the extraction. At first we have added this library into our java project. Then with the help of twitter app we have obtained Consumer Token Key and Access Token Key. Further, extraction of tweets will be start only after when we generate Access Key. Generation of Access Key needed every time for the extraction of the tweets. The twitter4j containing libraries are shown in Figure 5.1.

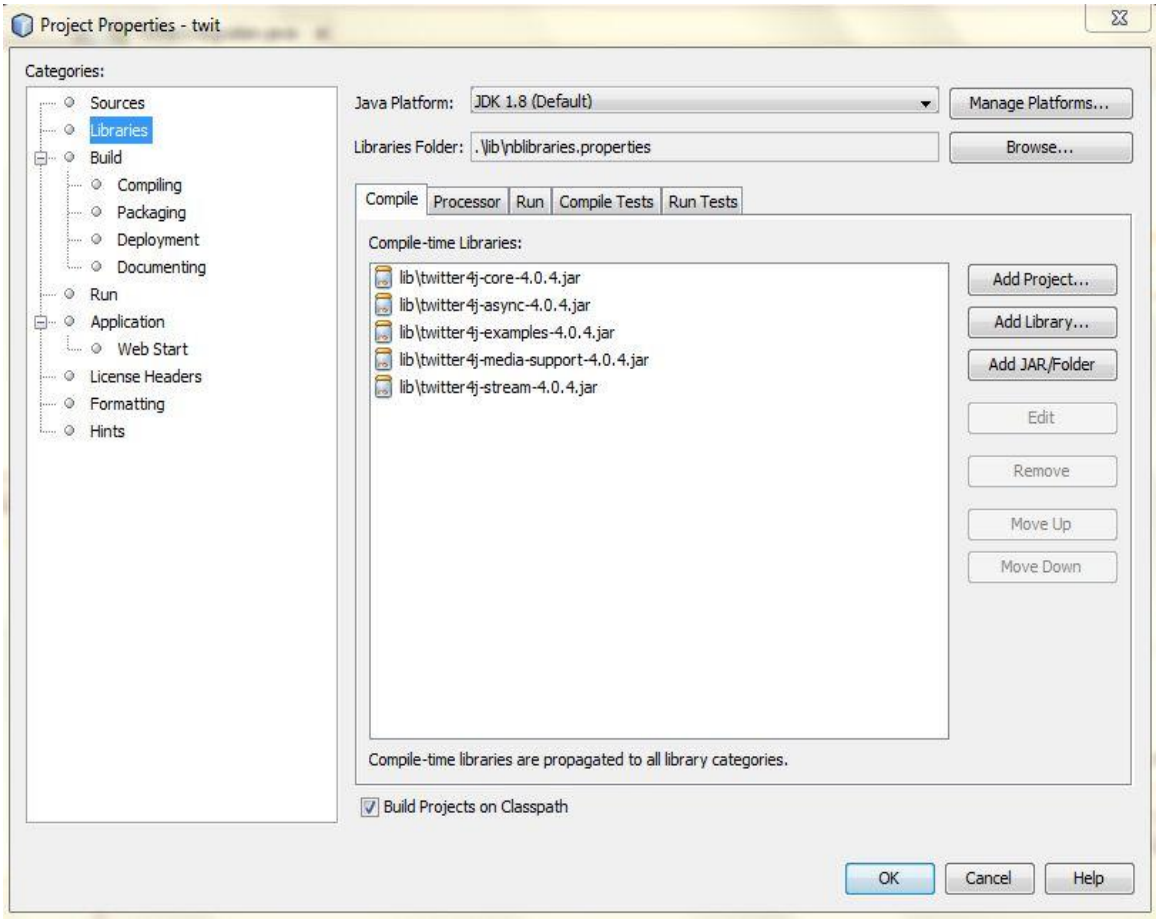


Figure 5.1: Twitter4j libraries

We have made twitter app to generate the consumer token key and access token key. Figure 5.2 shows the generation of consumer key and figure 5.3 shows the generation of access token key.

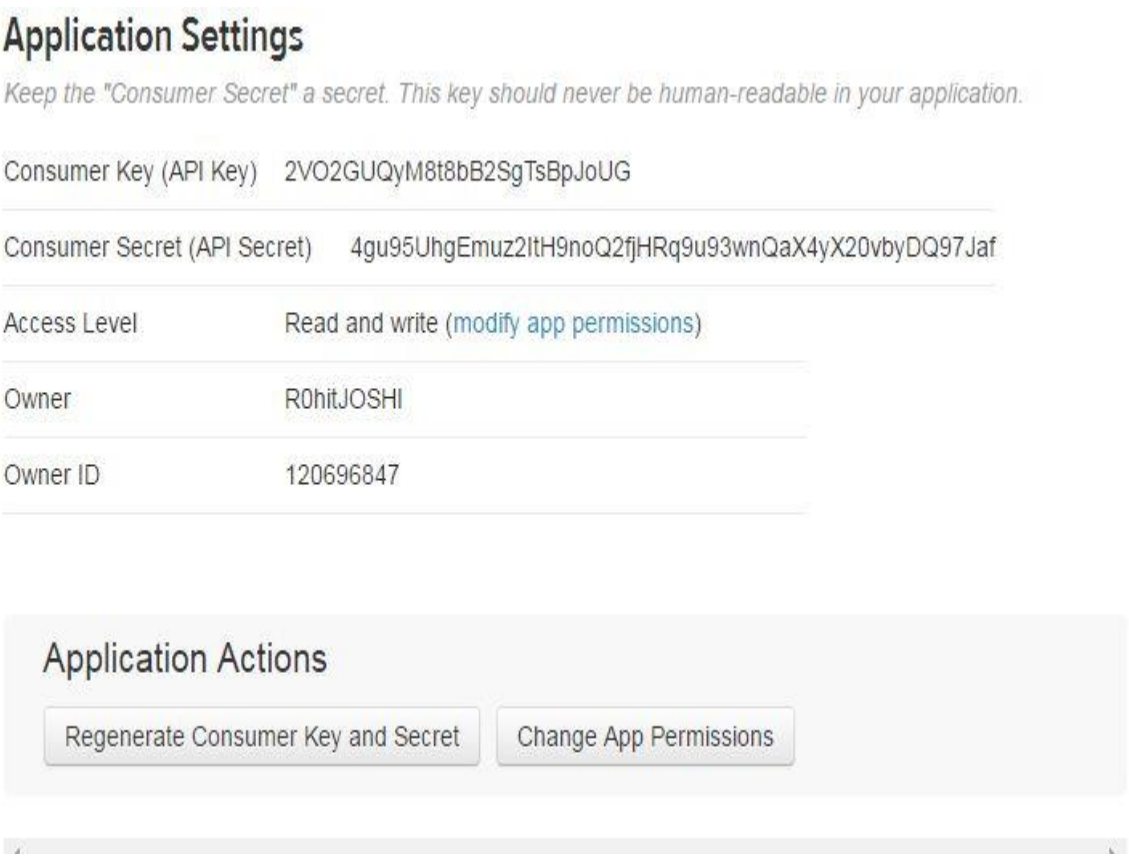


Figure 5.2 Consumer Token Key

Consumer Token Key will be provided by the twitter app. There is a unique key for every app and that key known as Consumer Token Key. In order to obtained tweets we have to apply consumer token key and access token key into the java code.

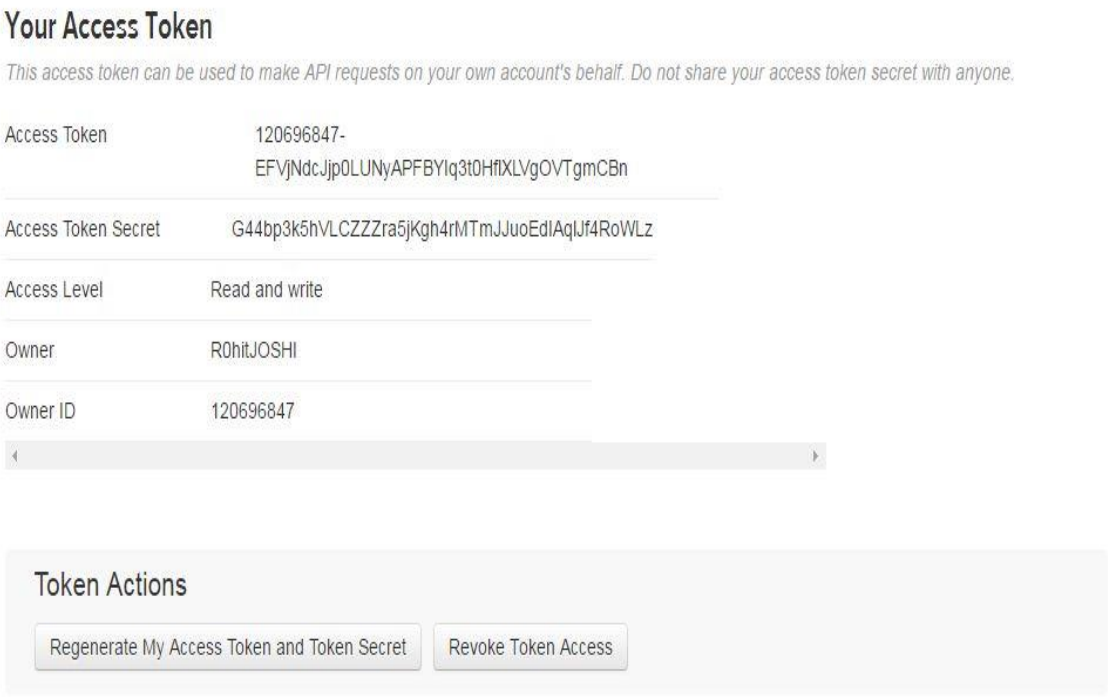


Figure 5.3 Generation of Access Token Key

8 22:15:37 IST 2016:#Budget2016. Thank you FM for withdrawing EPF tax proposal.
8 22:15:19 IST 2016:RT @dr_sandeepyadav: बीवी कह रही की सिगरेट के रेट बढ़ गए, अब तुम्हारी साल भर की सिगरेट के पैसे से मेरा एक डायमंड का सेट आ जाएगा छोड़ दो □...
8 22:13:57 IST 2016:Largely unnoticed in #Budget2016: 100% FDI in food marketing is game changer for farmers... https://t.co/u2tvNVGx5D
8 22:13:21 IST 2016:RT @ScotchWhiskySWA: #FairTaxforWhisky in #Budget2016 would recognise economic contribution of #Scotch. That's why w
8 22:12:53 IST 2016:RT @minhazmerchant: Breaking: FM Arun Jaitley withdraws #EPF tax from #Budget2016 proposal till "comprehensive revie
8 22:11:12 IST 2016:RT @pensionlawyers: View our latest video as @francoisbarker outlines the #pensions changes that may still be in sto
8 22:09:37 IST 2016:Senator Jay MOrris not buying what @stephenwags is hawking! #lalege #lagov #Budget2016
8 22:09:22 IST 2016:View our latest video as @francoisbarker outlines the #pensions changes that may still be in store in #Budget2016 ht
8 22:09:08 IST 2016:RT @pdputu42: शेयर बाज़ार में रिकवरी देश की आर्थिक हालत बाज़ार दर्शाता है मगू का खेत नही

8 22:08:11 IST 2016:RT @vidyarthee: Superb!! The #ModiGovt withdraws #EPF tax proposal under political & public pressure. #Congress fougl
8 22:07:26 IST 2016:RT @CAclubindia: Central govt withdraws EPF tax proposal : Finance Minister @arunjaitley #Budget2016
8 22:07:01 IST 2016:RT @IndiaSpend: Women & child development ministry gets 19% of #Budget2016 gender spend, 6% lower than 2014: https:/
8 22:06:27 IST 2016:RT @_RCCAO: .@markromoff discusses shovel ready vs. shovel worthy at #ocsoutlook16 @OntConstSec #Budget2016 https://
8 22:03:44 IST 2016:.@markromoff discusses shovel ready vs. shovel worthy at #ocsoutlook16 @OntConstSec #Budget2016 https://t.co/6BXtjJ8
8 22:17:26 IST 2016:RT @yogrishiramdev: आज लोकसभा के बजट सत्र में मोदी जी का मुस्कुराना व राहुल जी का गुस्साना, दोनों ही देखने को नहीं मिले

Figure 5.4 List of tweets

Figure 5.4 shows the list of tweets extracted from the Twitter4j API and these tweets can be converted to the excel files. These tweets contain user references, urls and punctuations.

5.1.2 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 labeled and stored in files as test dataset. We have two data sets: positive and negative. We have created two separate files for positive and negative set as shown in Figure 5.5 and Figure 5.6.

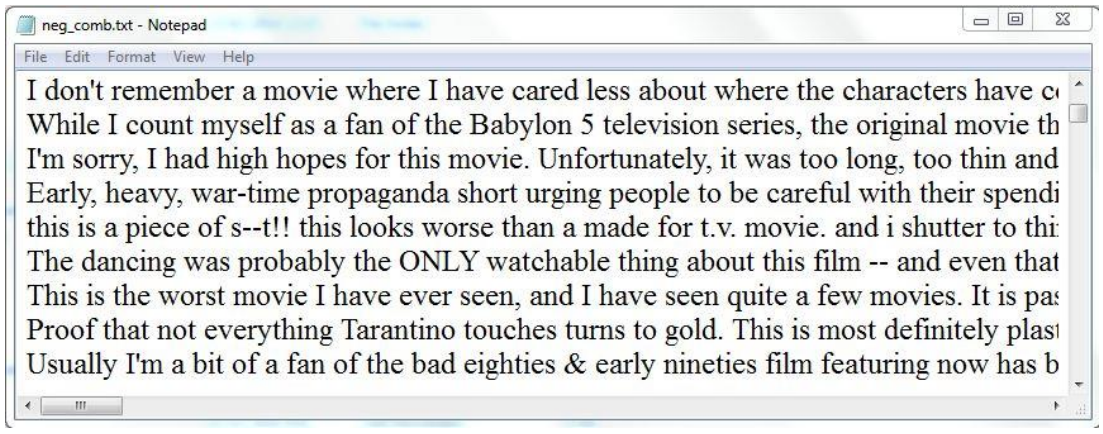


Figure 5.5 Positive Training Set

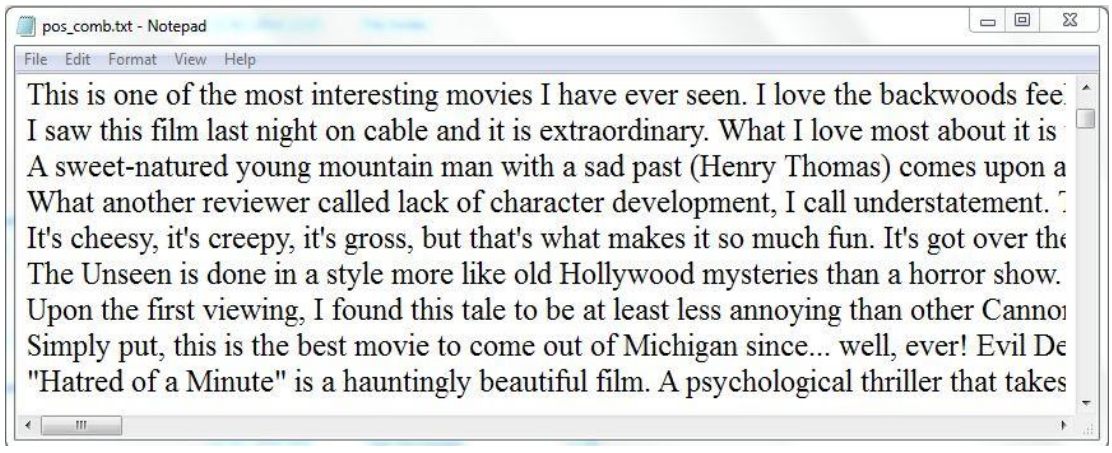


Figure 5.6 Negative Training Set

5.2 Results

We are using R language for implementation. R language offers maximum support when it comes machine learning techniques. Machine learning techniques can be easily implemented in R language. Packages that we are usimg are “RTextTool”,”Rweka”and “e1071”. RTextTools have most of the machine learning algorithms but not have Naïve Bayes, which is included in e1071 package and Rweka package is used for n-gram feature.

Table 5.1: Precision and recall for Unigram feature

Algorithm	Unigram	
	Precision	Recall
Naïve Bayes	0.75	0.71
Support Vector Machines	0.82	0.76
Maximum Entropy	0.74	0.70

Table 5.2: Precision and recall for Unigram feature

Algorithm	Bigram	
	Precision	Recall
Naïve Bayes	0.72	0.70
Support Vector Machines	0.76	0.71
Maximum Entropy	0.73	0.70

Table 5.3: Precision and recall for Hybrid feature

Algorithm	Hybrid	
	Precision	Recall
Naïve Bayes	0.73	0.71
Support Vector Machines	0.83	0.74
Maximum Entropy	0.76	0.73

The reason we are using R language because when the dataset is big, it is fast and efficient in terms of performing. The packages in the R tool are updated regularly and have greater number of probabilistic and statistical functions.

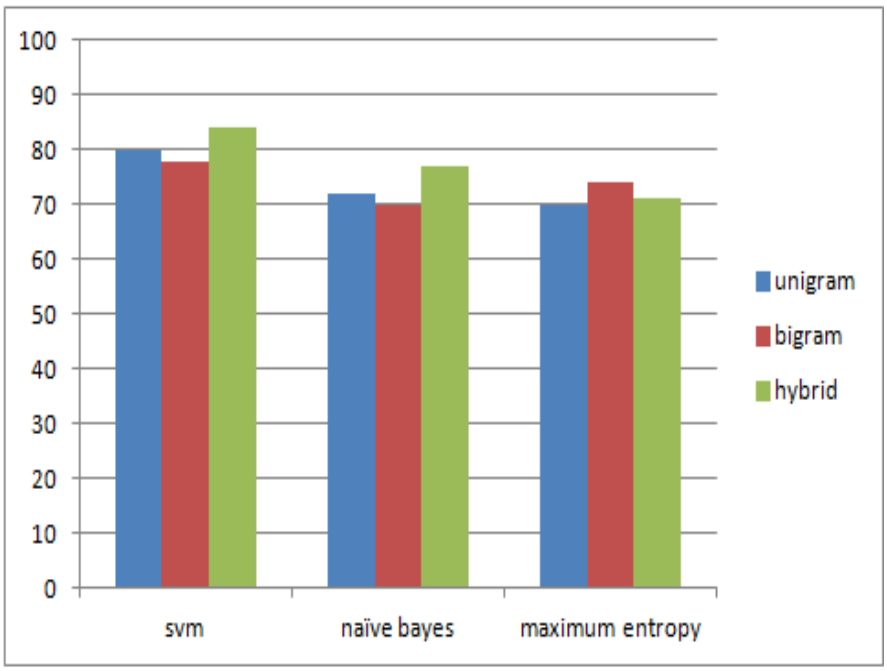


Figure 5.7 Results of machine learning algorithms

We have obtained the result as hybrid feature with svm classifier gives the best results for perediction of sentiment of twitter data. We obtained 84% accuracy using hybrid feature on svm classified data. 70% is least we have obtained in bigram with Naïve Bayes classifier. Max Ent outcomes Naïve Bayes in bigram feature and thus obtained 74% accuracy. The results can be shown in Figure 5.7.

Chapter 6: Conclusion and Future Scope

6.1 Conclusion

In this thesis, we have done comparative analysis on supervised classifiers like Naïve Bayes, support vector machines and maximum entropy using unigram, bigram and hybrid (unigram + bigram) feature . There is need to do sentiment analysis as texts in form of messages or posts to find the whether the sentiment is negative, positive or neutral. We had extracted data from twitter i.e. movie reviews for sentiment prediction using machine-learning algorithms. First we extracted the data from twitter using twitter API. Then in pre-processing, we clean the data and make the data available to train using classifiers. We have collected 15000 tweets for training set and 2000 tweets for testing set. SVM using hybrid feature outperforms all other classifiers and selection feature with accuracy of 84% .Max Ent surpass Naïve Bayes with bigram feature. MaxEnt, on some data sets gives better results than Naïve Bayes. It is concluded that SVM gives better results than other classifiers.

6.2 Future Scope

In future, we are planning to make automatic sentiment classifier for more than one languages starting from the Hindi language. As nowadays multilingual messages are posted in twitter, so we will able to predict the sentiment for any language.

References

[1] B. 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.

[2] L. Jiang, M. Yu, M. Zhou, X. Liu and T. Zhao, "Target-dependent twitter sentiment classification", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 151--160, 2011.

[3] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou and P. Li, "User-level sentiment analysis incorporating social networks", *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1397--1405, 2011.

[4] 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.

[5] M. Anjaria and R. Guddeti, "Influence factor based opinion mining of Twitter data using supervised learning", *Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on*, pp. 1--8, 2014.

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

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

[8] M. Koppel and J. Schler, "THE IMPORTANCE OF NEUTRAL EXAMPLES FOR LEARNING SENTIMENT", *Computational Intell*, vol. 22, no. 2, pp. 100-109, 2006.

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

[10]A. Andrew, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods20016Nello Christianini and John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 2000. xiii + 189 pp., ISBN: ISBN 0-521-78019-5 Hardback: £27.50", *Kybernetes*, vol. 30, no. 1, pp. 103-115, 2001.

[11]A. Berger, V. Pietra and S. Pietra, "A maximum entropy approach to natural language processing", *Computational linguistics*, vol. 22, no. 1, pp. 39--71, 1996.

[12]K. Nigam, J. Lafferty and A. McCallum, "Using maximum entropy for text classification", *IJCAI-99 workshop on machine learning for information filtering*, vol. 1, pp. 61--67, 1999.

[13]D. Romero, B. Meeder and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter", *Proceedings of the 20th international conference on World wide web*, pp. 695--704, 2011.

[14]S. Tan and J. Zhang, "An empirical study of sentiment analysis for Chinese documents", *Expert Systems with Applications*, vol. 34, no. 4, pp. 2622-2629, 2008.

[15]J. Martineau and T. Finin, "Delta TFIDF: An Improved Feature Space for Sentiment Analysis", *Proceedings of the Third International ICWSM Conference*, vol. 9, 2009.

[16]FA. Nielson. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs", *arXiv preprint arXiv:1103.2903*, 2009.

[17]SM. 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.

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

[19]K. Denecke, "Using SentiWordNet for Multilingual Sentiment Analysis", *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on IEEE*, pp. 507—512, 2008.

[20]B. Gokulkrishnan, P. Priyanthan, T. Ragavan, N. Prasath and A. Perara," Opinion Mining and Sentiment Analysis on a Twitter Data Stream", *Advances in ICT for Emerging Regions (ICTer), 2012 International Conference IEEE,* pp. 182—188, 2012.

[21]F. Neri, C. Aliprandi, F. Capeci, M. Cuadros and T. By," Sentiment Analysis on Social Media ", *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining,* pp.919—926, 2012.

[22]T. Wilson, J. Wiebe and P. Hoffman, "Recognizing Contextual Polarity: An Exploration of Features for Phrase-Level Sentiment Analysis", *Computational Lingusitics, MIT Press,* vol. 35, no. 3,pp. 399—433, 2009.

[23]N. Godbole, M. Srinivasaiah and S. Skiena," Large-Scale Sentiment Analysis for News and Blogs", *ICWSM,* vol. 7, no. 21, pp. 219—222, 2007.

[24]F. Benamara, C. Caserano, A. Picariello, DR. Recupero and VS. Subrahmanian," Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone", *ICWSM,* 2007.

[25]D. Boyd and N. Ellison," Social Network Sites: Definition, History, and Scholarship", *IEEE Engineering Management Review,* vol. 3, no. 38, pp. 16—31, 2010.

[26]A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment Analysis of Twitter Data", *Proceedings of the workshop on languages in social media, Association for Computational Linguistics,* pp. 30—38, 2011.

[27]T. Nasukawa and J. Yi, "Sentiment Analysis: Capturing Favorability Using Natural Language Processing", *Proceedings of the 2nd international conference on Knowledge capture, ACM,* pp. 70—77, 2003.

[28]H. Wang, D. Can, A. Kazemzadeh, F. Bar and S. Narayanan," A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle",

Proceedings of the ACL 2012 System Demonstrations, Association for Computational Linguistics, pp. 115—120, 2012.

[29]T. Wilson, J. Wiebe and P. Hoffman,"Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", *Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics,* pp. 347—354, 2005.

[30]H. Kanayam and T. Nasukawa, " Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis", *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,* pp. 355—363, 2006.

[31]Y. Choi and C. Cardie, "Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis", *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics,* pp. 793—801, 2008.

[32]P. Melville, W. Gryc and RD. Lawrence, " Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM,* pp. 1275—1284, 2009.

[33]G. Paltoglou and M. Thelwall, "A study of Information Retrieval weighting schemes for sentiment analysis", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics,* pp. 1386—1395, 2010.

[34]J. Fernandez, Y. Gutierrez, J. Gomez and P. Martinez-Barco, " GPLSI: Supervised Sentiment Analysis in Twitter using Skipgrams", *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), number SemEval,* pp. 294—299, 2014.

[35]T. Mullen and R. Malouf, "A Preliminary Investigation into Sentiment Analysis of

Informal Political Discourse", *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 159—162, 2006.

[36]A. Harb, M. Plantie, G. Dray, M. Roche, F. Trousset and P. Poncelet, "Web opinion mining: How to extract opinions from blogs?", *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*, *ACM*, pp. 211—217, 2008.

[37]SM. Kim and E. Hovy, "Determining the Sentiment of Opinions", *Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics*, 2004.

[38]A. Martalo, N. Novielli and F. De Rosis," Attitude Display in Dialogue Patterns",*Proc. AISB 2008 Symposium on Affective Language in Human and Machine*, 2008.

[39]E. 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, 2010.

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

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

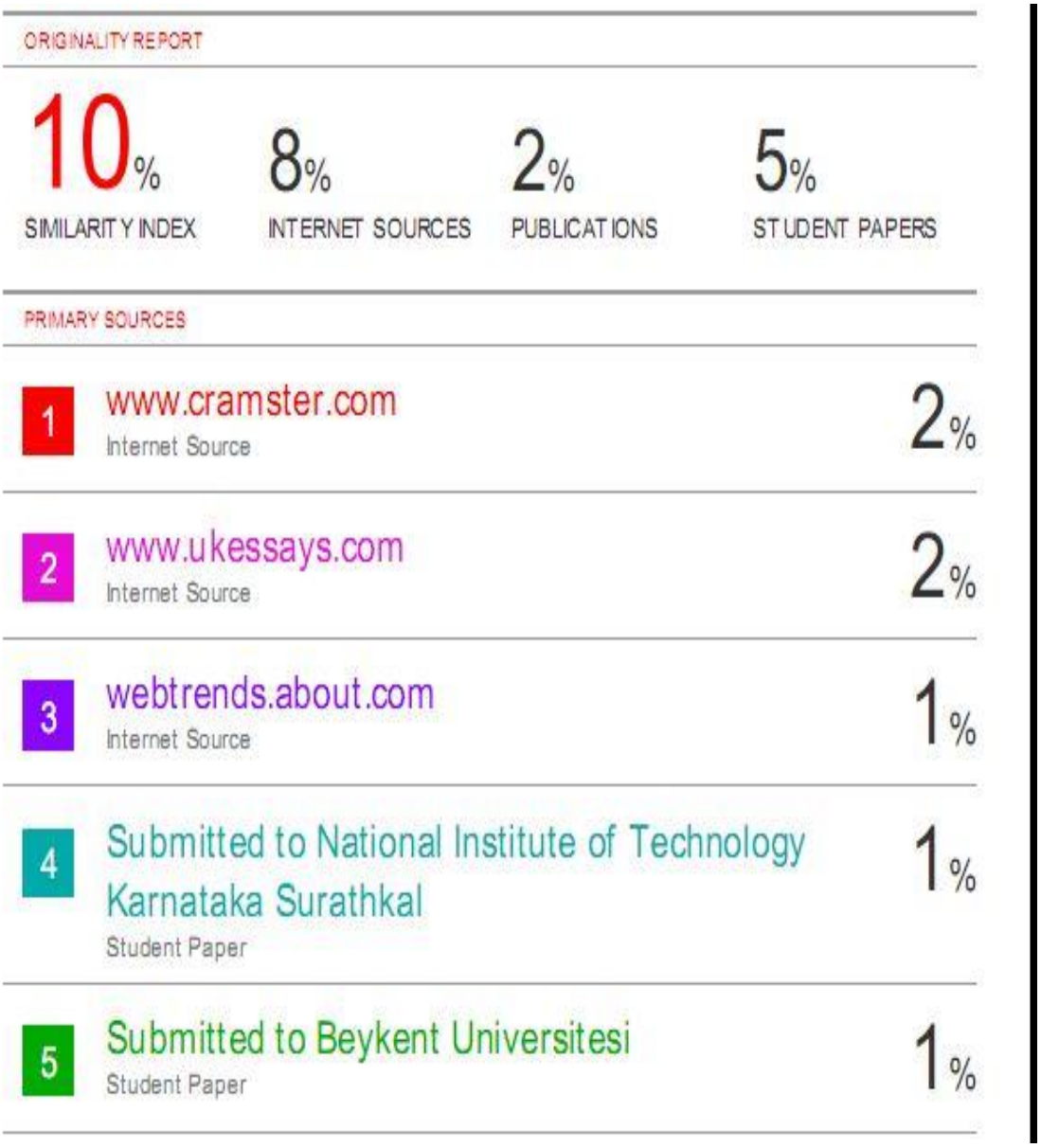
List of Publication

Rohit Joshi and Rajkumar Tekchandani, "Comparative analysis of twitter data using supervised classifiers", *in International Conference on Inventive Computation Technologies (ICICT 2016), IEEE*. [Accepted]

Video Link

https://youtu.be/0AO9wbO_v5c

Plagiarism Report



ORIGINALITY REPORT

10%	8%	2%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.cramster.com Internet Source	2%
2	www.ukessays.com Internet Source	2%
3	webtrends.about.com Internet Source	1%
4	Submitted to National Institute of Technology Karnataka Surathkal Student Paper	1%
5	Submitted to Beykent Universitesi Student Paper	1%