A Mixed Based Classifier Approach for Sentiment Analysis

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

Master of Engineering

in

Software Engineering

Submitted By

Sudhanshu Bhatia

Roll No. 801331028

Under the supervision of:

Dr. Ashutosh Mishra
Assistant Professor
CSED Department

Mr. Sumit Miglani
Assistant Professor
CSED Department

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR UNIVERSITY

PATIALA – 147004

June 2015

CERTIFICATE

I hereby certify that the work which is being presented in the thesis entitled, "*A Mixed Based Classifier Approach for Sentiment Analysis*", 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 *Dr. Ashutosh Mishra and Mr. Sumit Miglani* 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.

(Sudhanshu Bhatia)

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

(Dr. Ashutosh Mishra)

Assistant Professor

CSED Department

(Mr. Sumit Miglani)

Assistant Professor

CSED Department

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 would like to thank the Almighty, who has always guided me to work on the right path of the life.

This work would not have been possible without the encouragement and able guidance of my supervisor's **Dr Ashutosh Mishra** and **Mr. Sumit Miglani** I thank my supervisor's for their time, patience, discussions and valuable comments. Their enthusiasm and optimism made this experience both rewarding and enjoyable.

I am equally grateful to **Dr. Deepak Garg**, Associate Professor and Head, Computer Science & Engineering Department, a nice person, an excellent teacher and a well – credited researcher, who always encouraged me to keep going with work and always advised me with his invaluable suggestions.

I will be failing in my duty if I don't express my gratitude to **Dr. S.S. Bhatia,** Senior Professor and Dean of Academic Affairs, Thapar University, 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 whom I dearly miss and without whose blessings none of this would have been possible. To my parents, I own thanks for their wonderful love and encouragement. I would also like to thank my brother, since he insisted that I should do so. I would also like to thank my close friends for their constant support.

Date: July, 2015
Place: Thapar University, Patiala (Sudhanshu Bhatia)

ABSTRACT

The increasing expansion of social media stuff provides massive collection of textual information. People share their thoughts and views on the WEB. So sentiment analysis used to classifies the sentiments or the opinions from this huge amount of data. There are already many algorithms to find the sentiment form the data but there are many difficulties present to handle data like slang words and miss-spelling so the efficiency and the accuracy of these algorithms became poor. In this methodology the underlying idea is to achieve a particular accuracy rate by a new mixed algorithm by using different approaches like POS, N-Gram and some lexicon techniques.

Table of Content

List of Figures

Figure No. **Figure Name** **Page No.**

List of Tables

Table No. **Table Name** **Page No.**

Chapter 1

INTRODUCTION

1.1 Social Networking

Social networks are collection of networks that standardize institutions and groups in a self-ruled and unbiased manner, around a same goal. That is, dynamic and flexible model, with exemption and improvisation between the network, mutual trust and respect for individuality is must. Member of a social network collect and remove data, information and knowledge. There for highlight their importance is mandatory [2].

As RAPOPORT [3] showed, a social network is a platform of knowledge flow. So it is highlighting the main importance of a social network as it is through it that the real flow of knowledge between people, groups or institution with the same interest happens. Furthermore, it allows interdisciplinary connections when there is a common member in such networks. Interdisciplinary connection set up new concepts, communication between different areas [2].

"Social networking is the process of growing the number of one's business and social contacts by connecting through individuals". Now a day's social networking provides many facilities to its users like Facebook, YouTube, tweeter, LinkedIn, and lots of social networking sites provide messaging, video sharing, chat and many facilities.

Here are some social network prominent examples:

Twitter: Twitter is a free micro-blogging service. It allows to peoples to post their views on any social and unsocial topic. These posts are called 'Tweets'. Twitter users send their tweets and also share or re-tweet others post and a tweet can contain maximum 140 characters in a tweet [15].

Facebook: It is a free social networking website which allows to user to create their own Facebook profile and share their photos, videos and messages with their friends.

According a study peoples of united state spend more time on Facebook then any other social networking site [14].

YouTube: YouTube is a free video sharing website on which you can share your video with whole world. And anyone can access and download your videos. If you create an account on YouTube and share your video and if people likes your video and subscribe you channel the YouTube pay you for this and this is best site to show your talent to the whole world [14].

Instagram: It is an free online mobile application for photos and videos sharing and social networking services that enables their users to take pictures and videos and share on social networking platform such as Twitter, Facebook and Flicker [14].

Figure 1.1: Social networking roles

Online social networking like twitter is very popular platforms for communication for the peoples to logs thoughts, sentiments and opinion about anything from social events to daily chatter. Twitter is a micro blogging site having more than 250 million active user, these 250 million peoples post 400 million tweets/day. Allowing researchers and companies to gather and analyze data at scale. As a result, a big

number of studies have monitored the trending topics, memes, and notable events on OSNs, including political events, stock market fluctuations, disease epidemics, and natural disasters.

1.2 Sentiment Analysis

Sentiment analysis is referring to the use of Text analysis, Computational linguistics and NLP to recognize and draw out some relevant information from the source material. In present time WWW (World Wide Web) is transforming in to the interactive ecosystem that allow bi-directional 0communication. It is now running through blogs, social networking sites like Facebook, Twitter, Yelp, Amazon, online discussion form and many more platforms provided by WEB 2.0. The internet was meant to make the world a smaller place. People nowadays change their way to explore the views in the age of internet. People depends on each other like when someone wants to purchase any product, they will look its online reviews on e-commerce websites like Flipkart, Ebay, Snapdeal etc. and then make a decision [1]. Sentiment analysis is a research field where the goal is to recognize the opinions, emotions and subjectivity expressed in text document. Many users on the web are generating data too large for a normal user to analyze. So to brutalize this, we have various sentiment analysis techniques.

Sentiment analysis is a computational study of people's attitude, opinion and emotion towards the entity. Sentiment analysis is a method to find the opinion or the sentiment value of any text. The value can be found by applying various techniques and algorithms already created. The two words sentiment analysis and opinion mining are interchangeable. But some researchers said that sentiment analysis and opinion mining have some different notations. In Sentiment analysis we first identify the sentiment value from the text and then analyze it while in opinion mining we analyze the opinion of people and then analyze it. So target of sentiment analysis is to find opinion.

In sentiment analysis Knowledge base technique and machine technique are mainly used. In knowledge base approach requires complex structured or unstructured information of predefined emotions and trained data file to identifying sentiments [1]. Sentiment can be captured from various levels of unpurified data: at the level of sentence, paragraph, document, or clause [2]. Multiple techniques can be used for

identify different types of sentiment regardless of the level at which sentiment is taken. Machine learning uses a training data set to develop a sentiment classifier which is used to classify the sentiments. Using predefine dataset for entire emotions is simpler than Knowledge base approach. In this methodology, we use machine learning technique for classifying tweets.

Sentiment analysis on twitter data is quite difficult because length of the tweets is too short, stress words are used like "future" can be written as "fuuutttuuuurrrre", emoticons are used, and presence of slang words and emoticons also force to use a preprocessing step before feature extraction. Idea used in this proposed methodology is that first we extract the specific feature from and then remove those features from tweets to get a normal text. Using different data structure technique and different machine learning algorithm we used in this work.

It has been composed quantity of 20000 text posts from Twitter evenly split automatically between three sets of texts:

- Text containing slang words, misspelling and stress words.
- Texts containing negative and positive emotions, such as sadness, anger or disappointment and happiness, joy, satisfaction, pleasure respectively.
- Objective texts that only state a fact or do not express any emotion which is neutral.

This project executes semantic analysis of our corpus and we show how to build a sentiment classifier that uses the collected body as training data [3].

Exclusively, there exist two types of methods for sentiment analysis machine learning-based and lexical-based. Machine learning methods often rely on supervised classification approaches, where sentiment detection is framed as a binary. This approach requires label data to train the keywords. While one advantage of learning-based methods is, their ability to adapt and creating a training model for specified purposes and context, their drawback is the availability of labeled data and hence the low applicability of the method of new data. This is because labeling data might be costly or not even necessary for some tasks.

The simplest to detect the way polarity (i.e., positive and negative affect) of a message is based on the emoticons it contains. Emoticons have become popular in recent years, to the extent that some (e.g. <3) are now included in English Oxford

Dictionary. Emoticons are primarily face-based and represent happy or sad feelings, although a wide range of non-facial variations exist: for instance, <3 represent a heart and expresses love or affection. To extract polarity from emoticons, we utilize a set of common emoticons from as listed in Table l given below. This table also includes the popular variations that express the primary polarities of positive, negative, and neutral. Messages with more than one emoticon were associated to the polarity of the first emoticon that appeared in the text, although we encountered only a small number of such cases in the data. A recent work has identified that this rate is less than l0%. Therefore, emoticons have been often used in combination with other techniques for building a training dataset in supervised machine learning techniques.

Emoticons	Polarity	Symbols
	Positive	:-) or :), :-) or :), ;-) or ;), :-O or :o, :-P or :p, (H) or (h), (A) or (a), <:o)
	Negative	:-(or :(, :-S or :s, :'(, :-$ or :$, +o(, :^), |-), (U) or (u)
	Neutral	:-| or :|, >.<, :o, -_-, -.-, -_-', :x, :-x

Table l: Emoticon polarity and symbols

Sentiment analysis is an application which classifies the given text due to having positive, negative and in some case neutral. In the last 5 years companies are willing to notice employees and customer's thoughts for their work and management and employee's thoughts for their company. In this case, the concept of emotion and opinions instead of the large common part they have, are frequently analyzed in order to determine the strength of the opinions. Opinions can be closely related to the certain emotion's intensities such as happiness, love, anger, fear, surprise, and sadness and their subdivided tertiary and secondary. Many of the researchers and companies are releasing a product to fulfill the increasing demands for such kinds of tools, many of them claim to perform sentiment analysis of any type of document in every domain unfortunately there is nothing in the market that can analysis the sentiment of any kind of document.

The main reason behind sentiment analysis is so confusing that is word often take different meaning and they also belong to different emotions depending on the

domain in which they are being used and this is also confusing that in which situation and in which emotions that sentence write for example, in tweet having keyword improved so here improve can be used for both negative and positive phases. The word “improved” is associated with positive comments but “improve” more often used in negative ones.

There are some methods which are already in use. Detail about the following like: sentence level sentiment, document opinion analysis and feature based sentiment analysis. In the beginning, document opinion analysis determines whether the text is negative, positive or neutral for an object and has a range of accuracy from 70 percent to 80 percent based on amount of human input and type of text. Secondly sentence level sentiment analysis once aims to identify whether the sentence is subjective (opinionated) or objective, then to classify a subjective sentence and determine text subjectivity and polarity but also what in particular the text author liked or disliked about the object, extract object features that are commented by extracting object features that are commented, determining orientation of opinions (positive/negative/neutral), grouping feature synonyms and produce a summary.

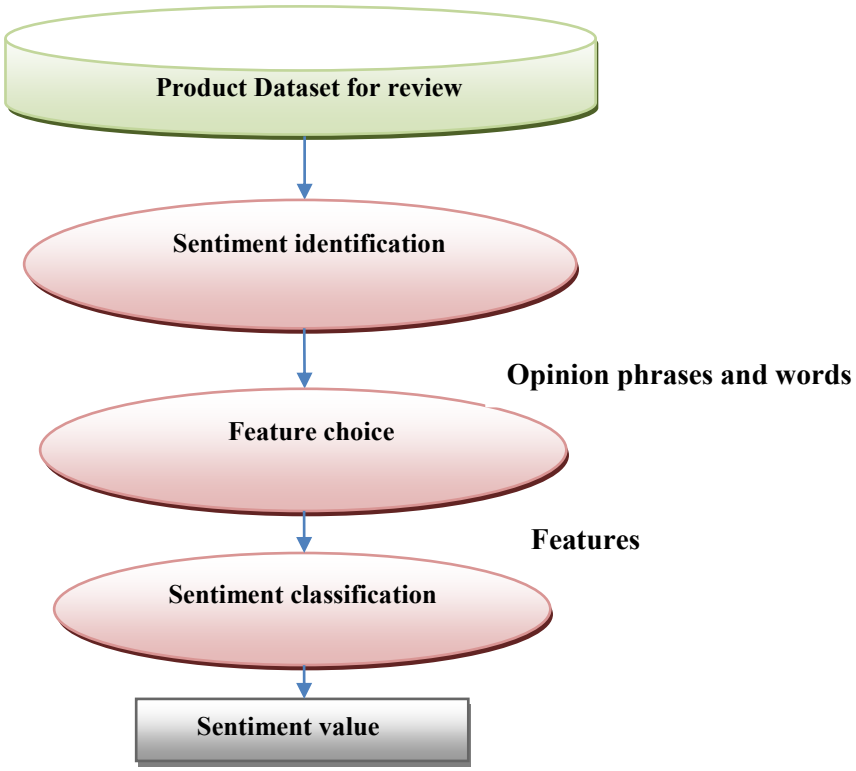


Figure 1.2: sentiment analysis process on product dataset review

Sentiment analysis is an interesting problem. Numbers of different approaches have been proposed in this research area. One popular approach for sentiment analysis is symbolic technique [6]. While another approach makes use of machine learning [8].

In adding up the technique for sentiment analysis, this work is also highlights some challenges and some important issues that need to be eliminate for sentiment analysis of tweets. These challenges generally have the disadvantage and advantage of different classifiers, adaptability to twitter traditional features, language used and also relation between structural properties of networks [7].

In this model, we focus on classifying positive and negative sentiments on document for general domains in conjunction with topic sentiment analysis based on the N-gram and Naïve-Bayes classifier and some recommender system for spell check. Naïve-Bayes works well for certain problems with highly dependent features This is surprising as the basic assumption of Naïve Bayes is that the features are independent, so we use N–gram technique in this methodology [5].

1.3 Things about Big data

Big data is the broad term for the dataset. A collective set of information in very large amount refer to big data. The traditional approaches for processing this data are failed. Some challenges with big data are: analysis, storage, sharing, information privacy and visualization. Big data is catching phrase or buzz word, used to show a big volume of both unstructured and structured data which is so large and difficult to process using traditional software techniques and databases.

META group analyst defined data growth challenges in three dimensional ways:

- Volume
- Velocity
- Variety

Volume: The first dimension to define the bigdata is volume refers to quantity of the information or data that is generated. Bigdata inferred vast volumes of data. The data in today's world is created by networks, machine and human interaction on social network sites like Facebook, WhatsApp and many other messengers and posting their views on blogs like twitter the volume of the data to be analyze is massive.

Velocity: The second aspect of Bigdata is velocity. Bigdata velocity deal with the amount of data generation in particular time unit, it can be say the pace of the data flows from sources like machines, business processes and interaction between human on social networking sites etc. The flow of data is large and regular. And we cannot handle this factor with real time solution.

Variety: The term 'variety' refers to the type of the data. That means which category the dataset belongs to is also an essential fact that should be known to data analyst. We store data like sources like spreadsheets and databases. Now data comes in the form of photos, videos, emails, pdf, monitoring devices and audio, etc. this variety of bigdata is difficult to store without knowing their variety.

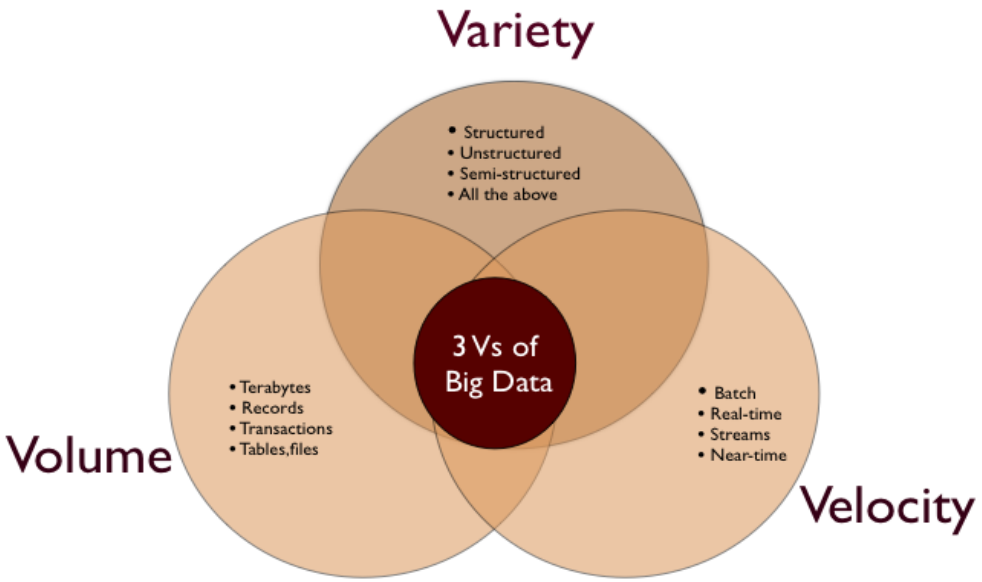


Figure 1.3: Three V's of big data and their constraints

As shown in figure 1.3, Big data is high volume, high velocity and/or high variety information assets that requires new form of analysis and processing that enables the enhanced decision making, process optimization and insight discovery. Twitter is having big data collection because millions of people are using twitter to share their views. And in our project we access these data and analyze if through some techniques.

1.4 Techniques used in Sentiment Analysis

Sentiment analysis techniques are divided in two three part which are Machine learning techniques, Lexicon based techniques and Hybrid technique [16]. Machine learning use linguistic feature and applies famous machine learning algorithm to classify the sentiment. Lexicon based technique divided in to corpus based or dictionary based approach which uses semantic or statistical method to find the sentiment polarity. Finally the hybrid approach is a mixed based approach with both machine learning and lexicon based. Machine learning and Lexicon based approach further divided in two many parts based on their types shown in the figure 1.4 below:

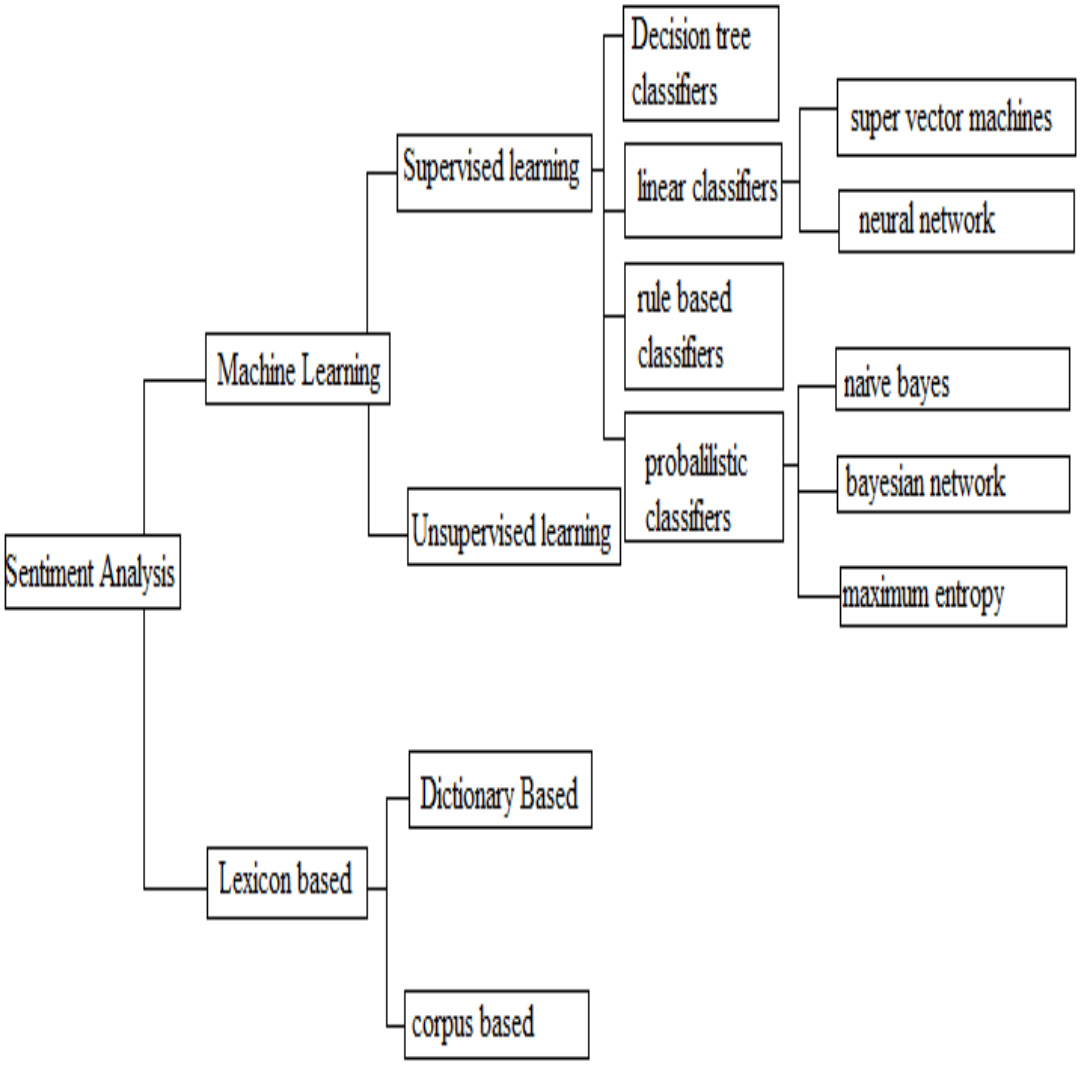


Figure 1.4: Sentiment analysis techniques

1.4.1 Machine Learning Technique

Machine learning technique is based on the famous machine learning algorithms that finds the sentiment analysis for a given text as Text Classification Problem (TCP). TCP make use of linguistic and syntactic feature.

TCP definition: set of n training records T = {C_1, C_2, C_3…….C_n} in the set 'T' all the records are labeled to a class. Classification model is related to feature in the record that belongs to a class labels. If any instance not belongs to any class then the model used to predict the class label for that particular instance. The soft classification is when all probabilistic value of labels assigned to class and hard classification is when only one instance label assigned to a class.

1.4.1.1 Supervised Learning

Supervised learning is a very common technique for classification problems. It mainly depends on the class labeled training document. It is most commonly used to training the decision tree and neural networks [17]. There are four classifiers used

- Decision Tree classifiers
- Linear classifiers
- Rule Based classifiers
- Probabilistic classifiers

1.4.1.2 Unsupervised Learning

The main role of classification is to classify all documents in to predefine category. Now the unsupervised learning approach also used by Guo and Xianghua [18] to find the aspect discussed in their local social review and sentiment expressed in different aspect. They find multi aspect global topics of social review and extracted sentiment based on a sliding window context from the text for this they have used LDA model.

Many unsupervised techniques that used semantic spaces, lexicon association using PMI [19] and to measure the similarity between polarity prototypes and words they used distribution similarity.

1.4.2 Lexicon Based Approach

In sentiment analysis positive sentiment express the wanted things and negative sentiment express the unwanted things. Opinion words are worked in sentiment analysis tasks. “Opinion lexicon” is some opinion idioms and phrases which are combine together. In the lexicon based we focused on the fixed expression that are using frequently in a text. To collect the opinion word list lexicon based have three approaches [17].

- Manual approach
- Dictionary based approach
- Corpus based approach

Manual check is not used frequently as alone because it is time consuming. Manual approach used with other dictionary and corpus based approach to check the final results generated by the automated approaches [17].

Chapter 2

LITERATURE SURVEY

In this chapter fundamentals of the learning techniques and understanding of work fulfilled in this report are describe. I will try to survey the literature on both twitter-specific and general sentiment analysis. But my main focus on those part which are more specific on this topic. First, I will try to explain the general sentiment analysis and then explain about twitter sentiment analysis. The work on sentiment analysis mainly focused on two things: first to identify whether a given data (text) is objective or subjective, and find the polarity of subjective data (text) [20].

Pang et al. [20] presents an Idea about sentiment analysis to represent the issue that makes sentiment analysis difficult more than other feature engineering, classification task and machine learning technique to sentiment analysis.

2.1 Work related to Pre-processing

Sentiment analysis has done in range of areas like sentiment analysis for product review [22], sentiment analysis study for blogs and news [23] and analysis on movie reviews [21]. Many researchers find that sentiment analysis is more difficult than traditional text classification. Number of classes in sentiment analysis is less than the classes in traditional classification [20]. In sentiment analysis, the text according to the classes can be positive or negative. The other multi valued classes are also there to show the text sentiment ‘positive’, ‘negative’ and ‘neutral’. In text classification technique we use the topic-based classification, which uses keywords to find the classification of any text that is why sentiment analysis is more difficult because in sentiment analysis do not work well with keywords.

2.2 Work related to Classifier

Number of researches has been done by many researchers in the sentiment analysis on twitter data. To analyze the sentiment from text there were so many techniques and there enhancement that were suggest in last few years. This survey gives a closer look to show these enhancements and to categorize and summarize some articles. There are two basic techniques to observe the sentiments form the twitter dataset.

- Symbolic technique

- Machine learning technique

2.2.1 Single Classifier Models

It used lexicon resources and Turney [4] used bag-of-word for sentiment analysis. In symbolic technique a text document is considered as a collection of words in document every word consider independent means there is no relation between any words in the document. Next step to determine the sentiment of every single word and these values are combined with some aggregation function (such as average or sum). Based on the average semantic orientation of tuples extracted from the review he found the polarity of a review. Tuples are phrases having adjective or adverb.

Kamps et al. [4] in his work used lexical database WN (WordNet) [5], which determined the emotional contents of a word with divergent dimensions. They determine the semantic orientation of adjective on the bases of distance metric developed on WN. WN database have words connected by synonym relation.

Balahur et al. [7] proposed EN (EmotiNet), it is conceptual presentation of dataset (text) that store the semantic of live events for a particular events of a specific domain and the text structure. Finite automata used in EN to identify the emotional response trigger by actions. Fine grained and coarse grained approaches are used to identify sentiment value in news headline. Binary classification of emotion used in coarse grained approach and classify emotions into different level in fine grained approach by one contributor of Semeval 2007 [8].

SWN (SentiWordNet) [1] support sentiment analysis application. SWN is a lexical technique, developed to brace the sentiment analysis. It gives an annotation based on three sentiment classes: negative, positive and neutral for each WN synset [9]. This lexical technique provides a synset sentiment representation.

Baroni et al. [6] proposed a system using word space model that reduces the complexities in lexical substitution tasks. It represents overall distributions along with the local context of a word.

Due to the requirement of a huge lexical database the knowledge base approach is found too difficult. Every second social network generated huge amount of data every day. Sometimes larger than the size of available lexical database, sentiment analysis became boring and incorrect. One of the scoring functions that classify the text defined as:

$$S_r = \frac{P(positive\_words \mid topic, r)}{P(negative\_words \mid topic, r)} = \frac{count(positive\_words \wedge topic)}{count(negative\_words \wedge topic)} \quad (1)$$

The sentiment sum S_r can be calculated as the ratio of the positive words to the negative words. Symbolic techniques are used to classify text sentiment value. This approach do not use training dataset, and classifiers attempts to classify text on the number of negative and positive word presents.

Words which indicate opinions are called 'opinion words' and lexicon is known as 'opinion lexicon'. Opinion finder is a simple subjective based lexicon, which has been used for sentiment analysis to compute confidence [21].

2.2.2 Multiple Classifier Models

Involves training classifiers that identify the sentiments in a given text documents. Machine learning technique use a training set classification and test set classification. Classification model is develop by using training set and tries to classify the input feature vector into corresponding class labels. Training set contain input feature vector and labels find the feature form the text. Manual explanation of data puts limits on the resulting training sets in term of quantity and quality. By using active learning, we can overcome this drawback. For this, a small portion of annotated data is used in order to classify unknown data which is then incorporated into the training [9]. One other approach involves to finding self-defining elements within the data in order automatically form the training set. Elements like keywords and emoticons have been studied in the past [10] [11], give promising results.

In machine learning some of the features like N-gram, POS (part of speech), term presence, negation and term frequency are used for sentiment analysis. Sentiment orientation of words, sentences, phrases and documents can be finding with these features. Naïve-Bayes is uses Bayes theorem. Naïve-Bayes classifiers are simple probabilistic classifiers. Conditional probability for Naïve-Bayes classifiers is defined as:

$$P(F \mid c_j) = \prod_{i=1}^{n} P(f_i \mid c_j) \quad (2)$$

Formula of Navie-Bayes to find the probability of positive and negative:

$$Probability\left[\begin{matrix}\frac{occurence\ of\ word1}{Total\ number\ of\ words} \mid \frac{occurence\ of\ word2}{Total\ numbure\ of\ words} \\ \mid \cdots \frac{occurence\ of\ wordn}{total\ number\ of\ words}\end{matrix}\right] *$$

$$|P(\ Tweets_{n..10}\ is\ positive)|\ P(\ Tweets_{n..9}\ is\ positive)| \cdots |P(Tweets_{n..0}\ is\ positive)|| \quad (3)$$

$$Probability\left[\begin{matrix}\frac{occurence\ of\ word1}{Total\ number\ of\ words} \mid \frac{occurence\ of\ word2}{Total\ numbure\ of\ words} \\ \mid \cdots \frac{occurence\ of\ wordn}{total\ number\ of\ words}\end{matrix}\right] *$$

$$|P(\ Tweets_{n..10}\ is\ negative)|\ P(\ Tweets_{n..9}\ is\ negative)| \cdots |P(Tweets_{n..0}\ is\ negative)|| \quad (4)$$

Here 'F' is the feature vector and c_j is the class labels. Feature vectors are defined as F={f1,f2,f3,…..fn}. Relationships between features are not considered in Naïve-Bayes. So the relationships between POS (part of speech) tag, Naïve-Bayes do not use negation and emotional keywords.

Domingos et al. [11] initiate that for some problems with highly dependent features Naïve-Bayes works fine. The features are independent this is the elementary assumption in Naïve-Bayes. Zhen et al. [12] proposed a new technique using efficient approaches for feature selection, classification and weight computation. Bayesian algorithm is used in this new proposed technique. By using unique features and representative features all the weights of classifiers are adjusted. Unique feature is the information used to distinguish the classes and representative information helps to present a class. On the bases of weight they calculate the probability and improve the Bayesian algorithm.

Barbosa et al. [13] proposed a two step automatic sentiment analysis technique for classify tweets. To decrease the labeling effort in developing classifier they use a noisy training dataset. First, they classify all tweets into objective and subjective tweets than select the subjective tweets and these subjective tweets classified as negative and positive tweets. Celikyilmaz et al. [14] gives a method for normalizing noisy tweets. This is a pronunciation based word clustering (PBWC) method. In PBWC, words their pronunciation are same these word clustered together and assigned a common token. To assign similar token for HTML links, numbers, target optimization for normalization and user identifiers they used text processing

technique. After normalization, to recognize polarity lexicons author used probabilistic techniques. To reduce error rate they perform classification by using the boos-texter classifier.

Wu et al. [12] give an influence probability model for sentiment analysis. Any tweet that begins with @username or @username found in the body of a tweet is a re-tweet that represents an influenced action and it contributes to influence probability. If @username exists in the tweet then it can be consider as a influencing action and this will contribute to the influencing probabilities. They experimented that there is high co-relation between these probabilities.

Xia et al. [17] for sentiment classifications they used a ensemble framework. This framework is obtained by combining many classification technique and feature sets. To structure the ensemble framework this methodology used three types of base classifier and two types of feature set. Using POS and word relation these two feature set are created. As a base classifier Support vector machine, Naïve-Bayes and Maximum entropy are selected. To obtain good accuracy and sentiment classification they apply different ensemble techniques like weighted combination, fixed combination and meta-classifier combination.

Pak et al. [18] developed a twitter dataset using twitter API by automatically collecting tweets and interpret those by using emoticon. A sentiment classifier they develop based on the multinomial NBC that uses part of speech tags and N-gram as feature. In this method, the training set is not much efficient since training set contain only tweets having emoticon and there is a possibility of error since emotion of the tweets in the corpus are labeled only based on emoticon's polarity.

G. Gautam et al. [24] used semantic analysis and machine learning approach for sentiment analysis on twitter dataset. They have used Naïve- Bayes, Maximum entropy, SVM and semantic analysis to find the feature extraction from tweets and get better accuracy. Naïve-bayes used because it simple during classifying stage and training. To classify the document according to their right category NB compare the content with the dictionary of words [25].

$$T^* = \arg_t P_{NB}(t \mid d)$$

$$P_{NB}(t \mid d) = \frac{(P(t) \sum_{i=1}^{n} p(f_i \mid t)^{n_i(d)})}{P(d)} \quad (5)$$

Here T* are the classes assign to tweets d. n(d) represent the feature count shows as f in the equation-4 found in tweet d. P(t) and P(t | d) are the parameter obtained through max likelihood estimate which increment by one. After the classification and training the data they use semantic analysis. For semantic analysis they derived from WN database. In WN database every term associated with each other. In their research the key work is to use the document stored in database that contain terms and identify similarity with the words that users use in their sentence [31].

For example, the sentence "we are sad" here the word "sad" is an adjective get selected and then they compared the selected word with the stored feature vector for similar words. Suppose three words: "sorrow", "unhappy" and "depress" be liable to very similar to the selected word "sad". Now they replace "sad" with any of the three words and give a negative sentiment.

The A (accuracy) in this proposed method measured in percentage and they computed as:

$$A = \frac{t_p + t_n}{t_p + t_n + f_n + f_p} \quad (6)$$

The R (recall) ratio, recall negative (n) and recall positive (p) are computed as:

$$R(p) = \frac{t_p}{t_p + f_n} \quad (7)$$

$$R(n) = \frac{t_n}{f_p + t_n} \quad (8)$$

The P (precision) ratio, precision negative (n) and precision positive (p) are computed as:

$$P(p) = \frac{t_p}{t_p + f_p} \quad (9)$$

$$P(n) = \frac{t_n}{t_n + f_n} \quad (10)$$

2.3 Some Feature of POS (Part of Speech)

Context-based features: for any word the context consists of POS tag of previous word that is unigram, Combination of POS tags of last two words i.e. bigram, current word and the next word predicted by the language model by processing the given sequence of word [26].

Word features capture property of word being tagged [26]. It includes,

- Suffixes: if word suffix is similar to a given suffix.
- Digits: the word is complete numeric or has any digits with it.
- Special characters: Does word contain any special character like '_' within it.
- English word: Occasionally occurrence of English word in Hindi text should be handled.

Corpus based features: this feature rely o information extraction from corpus data. Which are:

- Corpus word belongs to a single tag
- All possible tag of current word, as seen in corpus data.
- Does word occurred with only singe tag in corpus data.
- All possible tags of next word, as seen in corpus data.

Yan Dang et al. [27] propose lexicon enhancement technique for sentiment classification for an product reviews. They used Stanford POS tagger to perform tagging. To determine the sentiment value of extracted adverbs, nouns and adjectives they used SentiWordNet (SWN). In SWN each word has multiple senses. They have used prior polarity formulas for its verb, adverbs and adjectives. The average of the three polarity value calculated as:

Formula to find the score S as follows [27]:

$$S(w = POS)_{\,j} = \left(\sum_{k \in SWN(w=POS \text{ and } polarity=j)} SWN\_S(k)_j / |synsets(W = POS)|\right) \quad (11)$$

Here Part of speech (POS) ∈ (verbs, adverbs and adjectives), j∈ (objective, positive and negative) and k belongs to the synonyms set for the given word. Results calculated for positive (p), negative (n) and objective (o) [27]:

Score ("Good"= adjective) $_p$ =0.11,

Score ("Good"= adjective) $_n$ = 0.64,

Score ("Good"= adjective) $_o$ =0.25,

Score ("Good"= adverb) $_p$ = 0.06,

Score ("Good"= adjective) $_n$ = 0.56,

Score ("Good"= adjective) $_o$ = 0.38

2.4 N-gram Technique for Sentiment Analysis

N-Gram model, for a sentence from word$_1$……. word$_n$ the observed probability is approximated as [28]:

$$P(word_1 \dots\dots word_n) \prod_{j=1}^{n} P(word_1 | word_2 \dots\dots word_{n-1})$$

$$\approx \prod_{j=1}^{n} P\left(word_{n-j} | word_{n-j-1} \dots\dots word_1\right) \quad (12)$$

Here, this is assumed that to observe the ith word which is word$_i$ probability in the context of the preceded i-1 word can be approximate by the preceded n-1 word. It is the nth order Markov property.

From N-Gram frequency counts the conditional probability calculated as [28]:

$$p(word_i | word_{n-1} \dots\dots word_i) \frac{count(word_{n-1} \dots . word_2, word_1)}{count(word_{n-i} \dots\dots . word_1)} \quad (13)$$

M.A. Cabanlit et al. [29], this research work area aims to optimized the test feature selection in sentiment analysis for product reviews on twitter with polarity lexicons. This work proposed prospective to improve the performance of N-Gram based classifications.

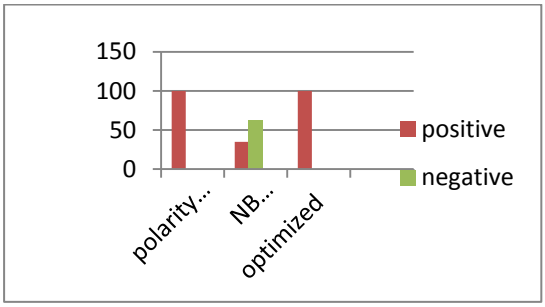


Figure 2.1:.Percentage of tweet classification for unigram

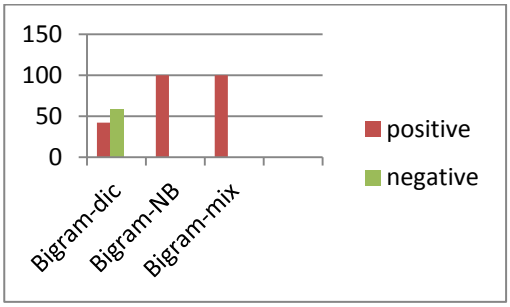


Figure 2.2: Percentage of tweet classification for Bigram

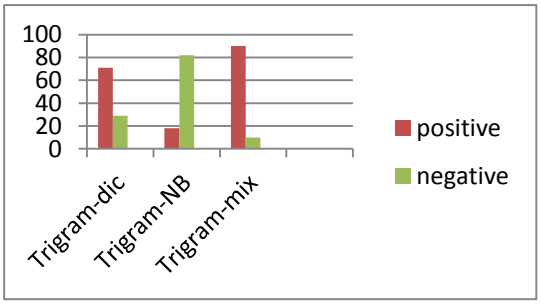


Figure 2.3: Percentage of tweet classification for Trigram

This study used a big training dataset and semantic analysis. They have used general dictionary for unigram and cross validate data for bigram and trigram [29]. They have used ten product reviews and tested how unigram, bigram and trigram would identify the positive and negative percentage. Graphs above depict percentage of negative and positive tweet [30].

Chapter 3

PROBLEM STATEMENT

In this particular chapter, the gaps which exist in the current work, problem statement of our proposed work, the objectives which are to be achieved and the method for achieving these objectives are discussed.

3.1 Gap Analysis

In the literature survey chapter, different steps involved in sentiment classification process have been discussed. Also, for each step, different techniques which can be used to create appropriate model. In the existing work following gaps exist:

1. Most of the existing techniques of sentiment Classifications are based on using single classifier.
2. Due to use of single classifier, inherent disadvantages of these classifiers affects the overall accuracy of the model.
3. Different classifiers can be combined, in serial (to improve the acceleration of the model), or parallel (to improve the accuracy).
4. Traditional Bayes classifier doesn't use current knowledge to calculate the posterior probability of the tokens.
5. Instead of using a single classifier for all users, multiple classifier can be used to analyze the tweets.

3.2 Problem Statement

Various sentiment analysis models have already been introduced for classification of twitter data based on certain features extracted from the tweets. There have been many improvements in the Naive Bayes classifier such as using N-gram with NB and different classifiers. Or applying different feature selection methods, pre-processing algorithms, to improve the accuracy of the classifier. But in every case, the main focus of the Naive Bayes model is to improve the maximum likelihoodness of the classifiers. As this model is based on frequency count, where on every occurrence of the token, either in good or spam mail, its frequency is increased by one. In order to overcome this problems, proposed model use local classifiers for each user and single global classifier.

3.3 Objectives

1. To study different techniques for each step of the Naive Bayes classifier.
2. To focus on improving sentiment classification accuracy of the model.
3. To use current knowledge of the bayes classifiers for calculating the increment done for each token, instead of increasing the value by one.
4. To test and validate the model by using different tweets corpus and comparing the result with the existing model.

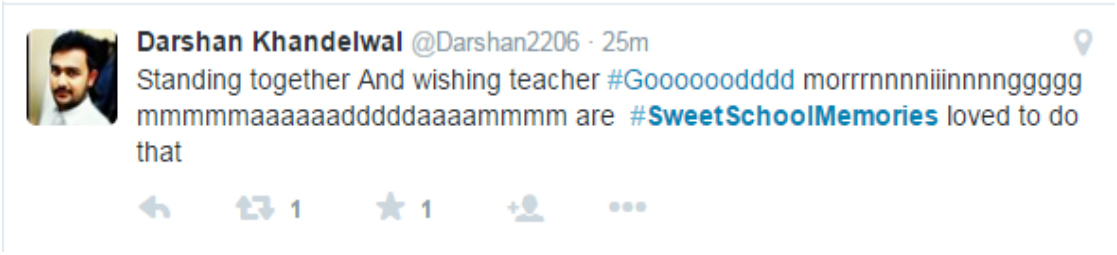
Chapter 4

METHODOLOGY

Sentiment analysis from twitter data is quite difficult because of short length of messages, misspelling, stress words and slang words. In our proposed solution we take dataset from twitter using twitter API and in this methodology sentence level sentiment analysis our work is done in three phases creation of dataset and dictionary for the sentiment analysis then removal of the # tags and @ form the tweets and also remove the re-tweets from dataset and finally do sentence level sentiment analysis on the data set.

4.1 Difficulties in Sentiment Analysis of Twitter Data

Sentiment analysis of the twitter data is very interesting task many algorithms are available to analyze the tweets. But to analyze the tweets here are some issues which are difficult to solve like, stop words, #tags, miss-spelling and most important stress words like shown in a tweet below:



Here shown in the tweet lots of stress words present. As known sentiment analysis is a result of a text to show that particular text is positive or negative but to analyze these kind of text is difficult.

4.2 Pre-processing

Using twitter API in this proposed methodology collect corpus of text. To store the datasets we use mongo database as NOSQL database for better performance. This proposed work having 20000 tweets in our dataset and a text file for positive, negative and a set of objective text (neutral sentiment). In the text file for positive and negative words and set the values for positive as 1, negative as 2 and for any neutral word added 3. This file contains 4818 negative, 4612 positive and 2000

neutral opinions (sentiment words) and there values. Emoticons are also classified based on their emotions like love, sadness, anger, joy. Table 2: represent the size of training set for keywords and emoticons.

Keywords		
Subjective-Objective	4000	4000
Positive-Negative	3500	3500
Anger-Sadness	1500	1500
Love-Joy	1500	1500

Emoticons		
Subjective-Objective	4000	4000
Positive-Negative	3500	3500

Table 2: Table shows Training dataset size

This methodology trained a classifier to recognize positive and negative sentiment. Text messages retrieved form twitter account, newspapers, magazine and individuals.

Stemming – Tweets are generally used with informal language and it include modern spelling and slang. Stemming will truncate words to their radical form so that play, playing, played, or plays will turn into 'score'. Think of stemming as linguistic normalization.

Sentiment and emoticons mapping – In twitter there are numerous emoticons which are used frequently in tweets. On the bases on used and mapped approach based on their emotions proposed work divide emoticons in positive and negative sentiment and discarded the emoticons that are inappropriate and ambiguous to sentiment.

4.3 Filtering of Tweets

- URLs do not intend to follow the short URLs and determine the content of the site, so regular expression matching or replace with generic word URL are used to eliminate all of these URLs.

- @username we eradicate"@username" via regex matching or replace it with generic word AT_USER.
- In Twitter text we have some #tags presents in the tweets so these #tags sometime having sentiment information so we just replace #tags with only tag values like #adidas replace with “adidas”.
- Punctuations and additional white spaces remove punctuation at the start and ending of the tweets. e.g. 'the college is very old!' replaced with 'the college is old'. And all the white spaces should replace with a single white space.

4.4 Proposed Algorithm for Feature Selection

This methodology successfully gets the feature from the dataset in an efficient way. This methodology use N-Gram and naïve-Bayes for the feature extraction and for data filtering we use some trained dictionary and Google translator for the miss-spelling. After tokenize all the tweets and normalize them this methodology use Google translator for all the miss spelling and got corrected first and then save them in to the library. To select the features from any text document many algorithms but this methodology use N-gram model for feature extraction. Suppose the sentence is “twitter is best for me” in this sentence or proposed model create many combinations for one-gram, bigram and so on like shown below in table …..

One-gram	Bigram	Trigram	Four-gram	Five-Gram
• Twitter • Is • Best • For • me	• Twitter is • Is best • Best for • For me	• Twitter is best • Is best for • Best for me	• Twitter is best for • Is best for me	• Twitter is best for me

4.5 Algorithm used for Sentiment Analysis

Algorithm 1 For training classifier the traditional Naive Bayes algorithm

1: Initialize each dataset entry as 0

2: **for** each Tweet from 1 to x **do**

3: **for** each token from 1 to y

4: Increase frequency count of y_i in class *t* by 1

5: **end for**

6: **end for**

7. $P(y_i|t) \leftarrow \frac{F(t,yi)}{\sum_{x_i} F(t,y_i)}$

In proposed model, ideally of increase frequency count by one, increment is done by posterior probability calculation across all users for that feature.

To calculating posterior probability formula is given below:

$$P^*(t \mid y) = \alpha P(t) \prod_{j=1}^{m} P(y_j \mid t) \tag{14}$$

Where, α is known as normalization factor, $P(y_j|t)$ is the local probability density, usually represented by CPT. $P(T = t)$ is the probability of class variable, i.e. prior probability. And $P(t \mid y)$ is the probability density of the feature learnt through each local classifier. It basically guides the amount of contribution added to the dataset, thus reflecting the confidence on current NB classifier.

Each conditional probability $P(y_j \mid t)$ in a dataset is calculated using the frequency count obtained from training data as:

$$P(y_j|t) = \frac{Frequency(y_j)}{\sum Frequency(y_j)} \tag{15}$$

Algorithm 2 To get opinion for a tweet

Input: Sentence S.

Output: opinion [Positive | Negative | Neutral]

1. Parse sentence S into S' using part-of-speech vector P to remove unnecessary words.
2. Tokenize sentence S' into a word vector W.

For each i=0 to Length (W):

- Generate all N-Grams up to Length (W) into new N-Gram vector G_i Using W.

3. **For** each j=0 to Length(G)
    - Analyze sentiments of G_j into result vector R_j.
4. Calculate opinion for sentence S using the result vector R

End.

In algorithm 2, input sentence S is first parsed into S' using Part of speech words, S' is formed by removing verbs, stop words, #tags and @username from S. On the other hand some words are also added to S' at the place of emoticons because emoticons can express the opinion about the tweet. After that new sentence S' is tokenized into a word list W, then we have generated all the possible N-Grams using the word list W. For example if word list W contains four words in that case the algorithm 1 will generate 4 unigrams, 3 bigrams, 2 trigrams and a fourgram and these resultant N-Grams are then stored into a new vector G. This then used to derive result vector R by analyzing sentiments of vector G. This proposed methodology used average of positive and negative sentiments stored into result vector R to calculate overall opinion about the original sentence S.

4.6 Classifying the Tweets

Traditional Naïve Bayes classifier is used in learning model, each tweet being classifies with individual classifier. In this proposed model, each token which is classifies with the help of Naïve Bayes and N-gram. This mixed based approach better accuracy and satisfactory results. Each N-gram after tokenization for the tweets, stemming algorithm is applied. For to storing the data proposed methodology used mongo database to store the tweet and their results. The overall impact of miss-spellings, stress words and stop words is cost minimum.

Figure 4.1 illustrate the approach for finding the sentiment score for user and followers. Four sections defined in the figure below are: data collection, getting user influence, getting the sentiment form tweet and result analysis. Dataset which is used in this methodology stores in mongo database as a .JSON file. Mongo database is used because to store a large file easily and give a better access then other databases. The corpus is having approximate 200000 tweets that is stored in the mongo database.

Implementation(cont'd.)

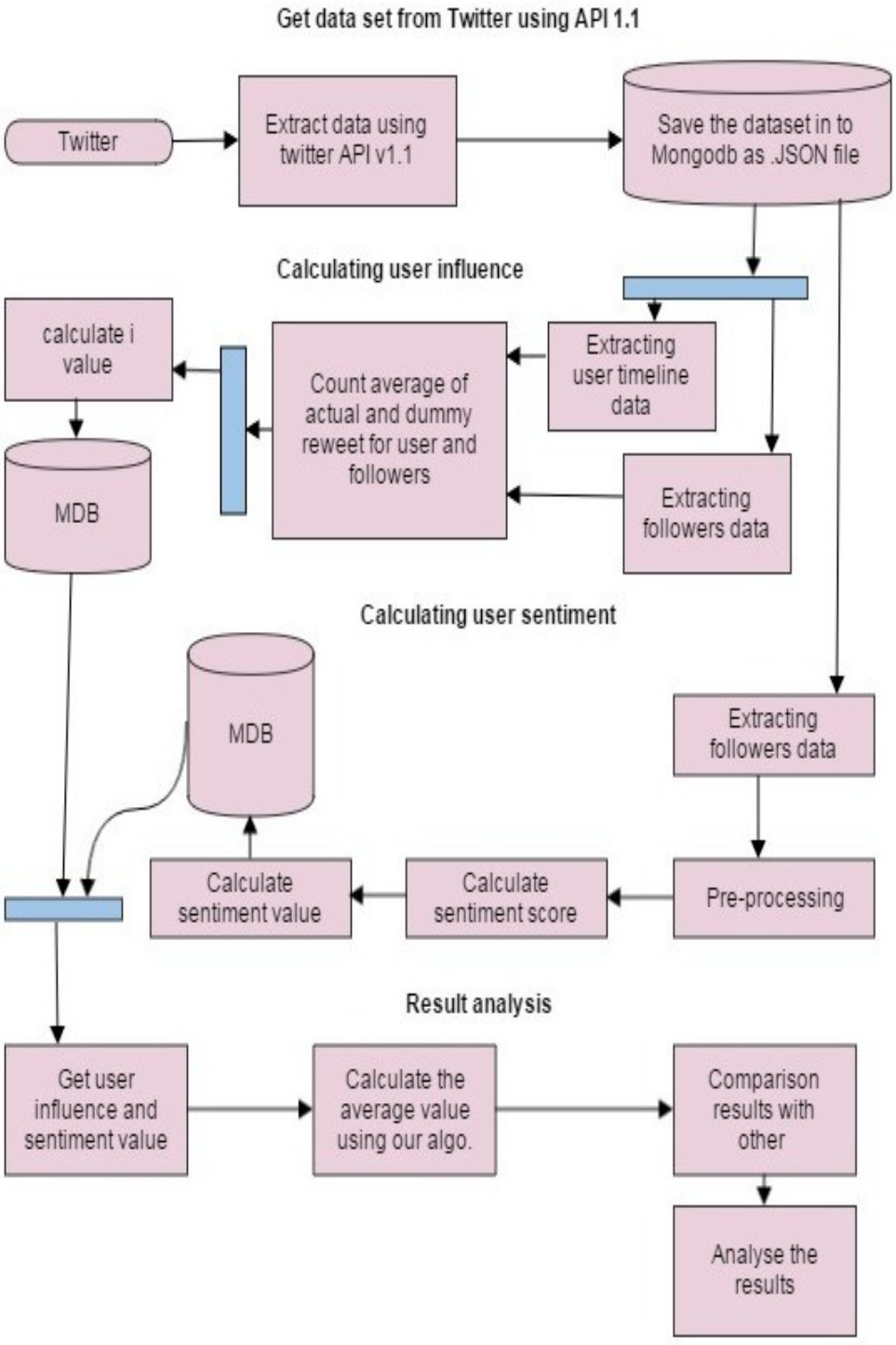


Figure 4.1: proposed model for sentiment analysis

Chapter 5

IMPLEMENTATION & RESULTS

This project is implemented on 64-bit windows 7 system having 4 GB of RAM. All the implementation work is done under Java framework. Radix encoded fragmented data structure is more efficient in place of hashmap data structure. This reduces searching time for the token in the database, which leads to an overall increase in the performance of the model.

5.1 Data used

In proposed model, Twitter dataset is used for testing. Twitter dataset consisting of tweets with different time. Nearly 20000 tweets, as .JSON file. For this experiment, First, the raw data is parsed for removing unwanted information and stopwords from the dataset. Then tokenization process is applied during which each token is assigned some attributes. This attributes control the spamity of a token, and are calculated using classifiers. With the introduction of each token, posterior probability of that token is calculated from all the local classifiers.

For calculation the difference between Naïve Bayes and proposed model, following evaluation measures are used. In the Table 5.1, $no_{h\to h}$ represents the number of Ham messages classified as Ham. Similarly $no_{h\to s}$ represents the Ham messages classified as Spam.

5.2 Implementation

We used NeBeans IDE 7.3.1 and mongo database 2.6.5 is used to process and store the tweet file. In the implementation of this project java version 1.8.0_20 is used to build this project. For this experiment we have used 20000 text collected from twitter streaming. We analyze sentence level sentiment using two predefined approaches and also analyze and find the result by our mixed approach. Our mixed-Based approach results are efficient and accurate then old approaches. Mixed based

approach use unigram, bigram, trigram for classifying the sentiment score.

5.2.1 Tokenize the Tweets

```
public class Token {

    String SplitString,tempString;
    String[] returnArray;
    ArrayList<String> temp,TokenizedText,nGram;
    int num,index;
    Boolean isIt;
    public Token(String SplitString) // Token(var_string).Tokenizer() or Token(var_string).tokenizerRemovePuncs()
    {
        this.SplitString=SplitString;

    }

    public ArrayList<String> Tokenizer()
    {
        int cont=index=0;
        TokenizedText= new ArrayList<String>();
        tempString="";
        for(int ctr=0;ctr<SplitString.length();ctr++)
        {
            while(Character.isLetterOrDigit(SplitString.charAt(ctr)) || SplitString.charAt(ctr)=='-'
             ||  SplitString.charAt(ctr)=='\'')
            {
                if(SplitString.charAt(ctr)=='-' || SplitString.charAt(ctr)=='3')
```

Figure 5.1: Token class

```
public ArrayList<String> tokenizerRemovePuncs()
{
    int cont=index=0;
    TokenizedText= new ArrayList<String>();
    tempString="";
    for(int ctr=0;ctr<SplitString.length();ctr++)
    {
        while(Character.isLetterOrDigit(SplitString.charAt(ctr)) || SplitString.charAt(ctr)=='-'
         || SplitString.charAt(ctr)=='\'')
        {
            if(SplitString.charAt(ctr)=='-' || SplitString.charAt(ctr)=='3')
            {
                ctr=isEmoticon(ctr);
            }
            tempString+=SplitString.charAt(ctr);
            ctr++;
            if(ctr>=SplitString.length())
                break;
        }
        if(tempString!="")
        {
            TokenizedText.add(tempString);
        }
```

Figure 5.2: To remove punctuations

```
    public int isEmoticon(int ctr2)
    {
        isIt=false;
        if(SplitString.charAt(ctr2)=='=')
        {

            if(SplitString.charAt(ctr2+1)==')')
            {
                tempString+=SplitString.charAt(ctr2);
                ctr2++;
                tempString+=SplitString.charAt(ctr2);
                ctr2++;
                if(ctr2<SplitString.length())
                {
                    while(SplitString.charAt(ctr2)==')')
                    {
                        tempString+=SplitString.charAt(ctr2);
                        ctr2++;
                        if(ctr2>=SplitString.length())
                            break;
                    }
                }
                isIt=true;
                TokenizedText.add(tempString);
                ctr2--;
                tempString="";

            } // =)
            else if(SplitString.charAt(ctr2+1)=='(')
```

Figure 5.3: If emoticon is present in the tweet

5.2.2 Implementation for Training the Dictionary

```
public class trainingCMD{

    static final String location = "D:/trainingData/";

    public static void main(String args[]){


        int again = 1;
        do{
            Scanner sc = new Scanner(System.in);
            System.out.println("Enter sentence: ");
            String input_sentence = sc.nextLine();
            Token tk = new Token(input_sentence);
            ArrayList<String> tkn = tk.tokenizerRemovePuncs();
            String []tokenized = tkn.toArray(new String[tkn.size()]);
            System.out.println(tkn);
            String input_gram = "";
            int i=0, polarity;
            for(i=0; i<tkn.size()-1; i++){
                input_gram = tokenized[i]+" "+tokenized[i+1];
                System.out.print(input_gram+" >>");
                polarity = sc.nextInt();
                writeToTextFile(input_gram,polarity);

            }

            for(i=0; i<tkn.size()-2; i++){
                input_gram = tokenized[i]+" "+tokenized[i+1]+" "+tokenized[i+2];
                System.out.print(input_gram+" >>");
                polarity = sc.nextInt();
```

Figure 5.4: Training the CMD Library

```
public static void writeToTextFile(String nGram, int x)
{



    if(x==1){ //positive
        writePos(nGram);

    }
    else if(x==2){//negative
        writeNeg(nGram);
    }
    else if(x==3){//neutral
        writeNeg(nGram);
        writePos(nGram);
    }



}


public static void writePos(String nGram){

    try
    {
        FileWriter fileTrain;
        BufferedWriter writer;
```

Figure 5.5: Training with using N-Gram

5.3 Results

For this experiment we have used 20000 text collected from twitter streaming. We analyze sentence level sentiment using two predefined approaches and also analyze and find the result by our mixed approach. Our mixed-Based approach results are efficient and accurate then old approaches. Mixed based approach use unigram, bigram, trigram for classifying the sentiment score.

Performance Measurement (%)	
Positive Recall	91.2
Negative Recall	86.7
Positive Precision	49.5
Negative Precision	38.8

Table 3: Naïve Bayes classification measurement

Performance Measurement (%)	
Positive Recall	90.2
Negative Recall	86.4
Positive Precision	47.3
Negative Precision	36.3

Table 4: Mixed based classification measurement

Table 1, shows the performance measurement using Naïve Bayes classification measurement with positive and negative recall and precision and Table 2, shows the performance measurement using Mixed-Based classification measurement.

Classifiers	**Max. Accuracy**	**Average Accuracy**	**Average F(*10)**
SVM	71.84%	68.98%	6.88
Naïve Bayes	84.96%	80.15%	8.09
Mixed algo(using N-Gram)	92.45%	91.18%	9.20

Table 5: Classification algorithm comparison

As shown in Table 3, for sentiment analysis we have already some algorithms like NB and SVM who are giving good accuracy and results. But with our approach we get better accuracy. Using N-Gram with symbolic technique we analyze 20000 tweets and compare it with other algorithm data.

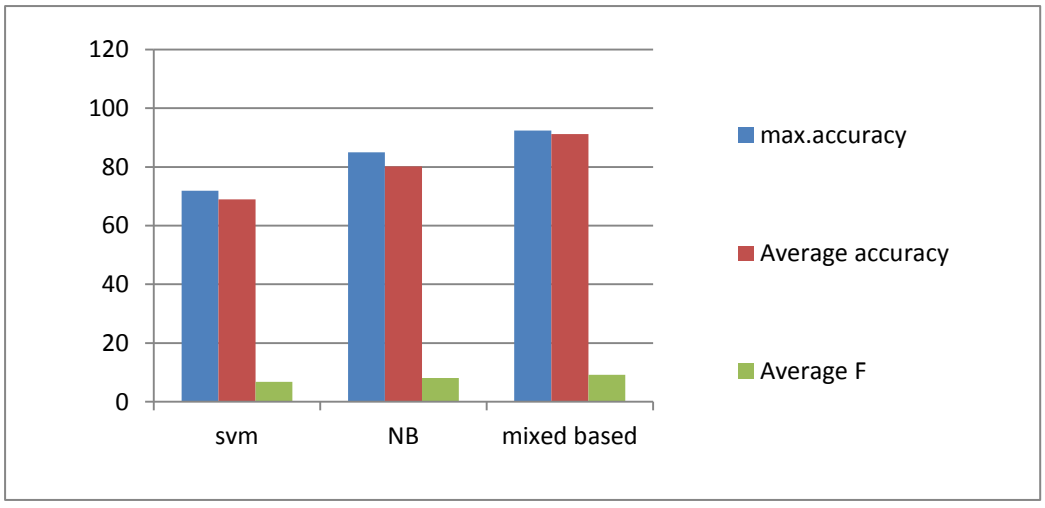


Figure 5.6: Comparison with NB and SVM

Tweet related to any particular topic are collected using Twitter API. Since there

is no need to analyze subjective and objective tweets separately. Performance of different sentiment classifiers are shown in the Figure 5.2 the graphical representation describes, naïve bayes having better accuracy rate then SVM which is 84.96% and 71.84% at maximum accuracy rate respectively but with our approach it is 92.45% and having better efficiency rate then these two. As we are using N-Gram approach it is N-Gram is 7.49% high accurate then NB (Naïve Bayes).

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

Thus, it has been proved that number of techniques can be used to perform the sentiment analysis. Each technique has its specific domain. Accuracy of the algorithm matters on the data be used for process if the source of the data is from social networking sites then the language and convention need to be address.

- Improved classification accuracy and efficiency of sentiment analysis of twitter data
- More emphasis on classification accuracy

- Sharing of knowledge among different users and use of current knowledge for improving performance.

6.2 Future Scope

Hence, future scope in sentiment classification domain is the accuracy and efficiency. Our proposed methodology as of now giving good accuracy but in many tweets Hindi words are also included which has no sentiment value so in future research can work on that and also will work on the data filtering to get the meaning of non English and stress words.

References

[1] M. S. Neethu and R. Rajasree "Sentiment Analysis in Twitter using Machine Learning Techniques" 4TH ICCNT, pp. 46-51, IEEE, 2013.

[2] C. L. Fink, D. S. Chou, J. J. Kopecky and A. J. Llorens " Coarse- and Fine- Gralned Sentiment Analysis of Social Media Text" JOHNS JOPKINS APL TECHNICAL DIGEST, vol. 30, number 1, 2011

[3] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.

[4] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Machine Learning, vol. 29, no. 2-3, pp. 103–130, 1997.

[5] Y. H. Cho and K. J. Lee, "Automatic affect recognition using natural language processing techniques and manually built affect lexicon," *IEICE - Trans. Inf. Syst.*, vol. E89-D, no. 12, pp. 2964–2971, Dec. 2006

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, ser. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86.

[7] Z. Niu, Z. Yin, and X. Kong, "Sentiment classification for microblog by machine learning," in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, 2012.

[8] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '03, 2003, pp. 105–112.

[9] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak, "Exploiting emoticons in sentiment analysis," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ser. SAC '13. New York, NY, USA: ACM, 2013, pp. 703–710.

[10] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *Proceedings of the ACL Student Research Workshop*, ser. ACLstudent '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 43–48.

[11] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," Machine Learning, vol. 29, no. 2-3, pp. 103–130, 1997.

[12] Y. Wu and F. Ren, "Learning sentimental influence in twitter," in Future Computer Sciences and Application (ICFCSA), 2011 International Conference on, pp. 119–122, IEEE, 2011.

[13] F.M.N. Hanif and G.A.P.Saptawati, “Correlation analysis of user influence and sentiment on twitter data” pp. 14-19, IEEE,2014

[14] B. Boyd and D. Ellison “social network sites: history and definition” from www.onlinelib.com/doi/10.111/j.1083-6101.2007.000393.x/abstract

[15] H. Saif, Y. He and H. Alani, “Semantic Sentiment Analysis of Twitter,” Proceedings of the 11th international SemanticWeb Conference, 2012.

[16] Diana Maynard, Adam Funk. Automatic detection of political opinions in tweets. In: Proceedings of the 8th international conference on the semantic web, ESWC’11; 2011. p. 88–99.

[17] W. Medhat, A. Hassan and H. Korashy, “Sentiment analysis algorithm and application: A Survey”, In asej, pp. 1093-1113, vol. 5, Elsevier 2014.

[18] F. Xianghua, L. Guo, G. Yanyan and Wang Zhiqiang, “Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon”, Knowl-Based Syst, 37, pp. 186–195, 2013.

[19] P. Turney, “semantic orientation applied to unsupervised classification of reviews”, Proceedings of annual meeting of the Association for Computational Linguistics (ACL’02); 2002.

[20] Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1_135, 2008. ISSN 1554-0669.

[21] Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classi_cation using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79_86.Association for Computational Linguistics, 2002.

[22] Dave, S. Lawrence, and D.M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classi_cation of product reviews. In Proceedings of the 12th international conference on World Wide Web, pages 519_528. ACM, 2003. ISBN 1581136803.

[23] Godbole, M. Srinivasaiah, and S. Skiena. Large-scale sentiment analysis for news and blogs. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM). Citeseer, 2007.

[24] G. Gautam and D. Yadav, “Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis”, contemporary computing (IC3), 2014 international conference, IEEE, pp. 437-442, 2014

[25] L.Liu, X.Nie,and H.Wang, ”Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis”, 5th Image International Congress on Signal Processing (CISP), pp. 1620 – 1624,2012.

[26] Dalal, A, Kumar, N, Sawant, U, Shelke, S.: Hindi part-of-speech tagging and chunking: A maximum entropy approach. In: Proceeding of the NLPAI Machine Learning Competition. (2006).

[27] Y. Dang, Y. Zhang and H. Chen, "A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews", IEEE Computer Socity, IEEE, vol. 25, pp. 46-53, 2010.

[28] Trim, C. (2013). What is Language Modeling? Retrieved from http://trimc-nlp.blogspot.com/2013/04/language-modeling.html

[29] M.A. Cabanlit and K.J. Espinosa, "Optimizing N-Gram Based Text Feature Selection in Sentiment Analysis for Commercial Products in Twitter through Polarity Lexicons" Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference, IEEE, pp. 94-97,2014.

[30] J. Wang and C. Zhang, "The Sentiment Trend Analysis of Twitter Based on Set Pair Contact Degree", IJCSI International Journal of Computer Science, 10,pp. 798-804, 2013.

[31] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis, Foundations and Trends in Information", Foundations and Trends in Information Retrieval, 2, 1-135, 2008.

[32]

List of Publications

Sudhanshu Bhatia, Dr. Ashutosh Mishra and Sumit Miglani.: A Mixed Based Classifier Approach for Sentiment Analysis: Communicated at Eighth International Conference on Contemporary Computing (IC3), IC3-2015, (2015)

Video link

https://www.youtube.com/watch?v=u-TtzHnTGZA