Sentiment Analysis Using Machine Learning Techniques

Abinash Tripathy

Department of Computer Science and Engineering
National Institute of Technology Rourkela

Sentiment Analysis Using Machine Learning Techniques

Dissertation submitted in partial fulfillment

of the requirements of the degree of

Doctor of Philosophy

in

Computer Science and Engineering

by

Abinash Tripathy

(Roll Number: 512cs1015)

based on research carried out

under the supervision of

Prof. Santanu Kumar Rath

May, 2017

Department of Computer Science and Engineering

National Institute of Technology Rourkela

Department of Computer Science and Engineering
National Institute of Technology Rourkela

May 19, 2017

Certificate of Examination

Roll Number: *512cs1015*
Name: *Abinash Tripathy*
Title of Dissertation: *Sentiment Analysis Using Machine Learning Techniques*

We the below signed, after checking the dissertation mentioned above and the official record book (s) of the student, hereby state our approval of the dissertation submitted in partial fulfillment of the requirements of the degree of *Doctor of Philosophy* in *Computer Science and Engineering* at *National Institute of Technology Rourkela*. We are satisfied with the volume, quality, correctness, and originality of the work.

Santanu Kumar Rath
Principal Supervisor

Ashok Kumar Turuk
Member, DSC

Pankaj Kumar Sa
Member, DSC

Bidyadhar Subudhi
Member, DSC

N. P. Gopalan
External Examiner

Sanjay Kumar Jena
Chairperson, DSC

Durga Prasad Mohapatra
Head of the Department

Department of Computer Science and Engineering
National Institute of Technology Rourkela

Prof. Santanu Kumar Rath
Professor

May 19, 2017

Supervisor's Certificate

This is to certify that the work presented in the dissertation entitled *Sentiment Analysis Using Machine Learning Techniques* submitted by *Abinash Tripathy*, Roll Number 512cs1015, is a record of original research carried out by him under my supervision and guidance in partial fulfillment of the requirements of the degree of *Doctor of Philosophy* in *Computer Science and Engineering*. Neither this dissertation nor any part of it has been submitted earlier for any degree or diploma to any institute or university in India or abroad.

_____
Santanu Kumar Rath

Dedication

I want to dedicate this thesis to my parents, wife and my beloved son Ribu.

Abinash Tripathy

Declaration of Originality

I, *Abinash Tripathy*, Roll Number *512cs1015* hereby declare that this dissertation entitled *Sentiment Analysis Using Machine Learning Techniques* presents my original work carried out as a doctoral student of NIT Rourkela and, to the best of my knowledge, contains no material previously published or written by another person, nor any material presented by me for the award of any degree or diploma of NIT Rourkela or any other institution. Any contribution made to this research by others, with whom I have worked at NIT Rourkela or elsewhere, is explicitly acknowledged in the dissertation. Works of other authors cited in this dissertation have been duly acknowledged under the sections ‘‘Reference’’. I have also submitted my original research records to the scrutiny committee for evaluation of my dissertation.

I am fully aware that in case of any non-compliance detected in future, the Senate of NIT Rourkela may withdraw the degree awarded to me on the basis of the present dissertation.

May 19, 2017
NIT Rourkela

Abinash Tripathy

Acknowledgment

First of all many thanks to God and my parents for showing me the proper way to success. I owe deep respect to all, who have helped me a lot and have contributed greatly for the completion of this thesis.

I would like to express my sincere gratitude to my supervisor, Prof. Santanu Kumar Rath for his guidance in the field sentiment analysis and also providing me a platform to work on this area. His profound insights and attention to details have been true inspirations to my research.

I am thankful to my Doctoral Scrutiny Committee (DSC) members, Prof. S. K. Jena, Prof. A. K. Turuk, and Prof. P. K. Sa of Computer Science and Engineering Department and Prof. B. Subudhi of Electrical Engineering Department for extending their valuable suggestions, and help whenever I approached them.

It is my great pleasure to show indebtedness to my friends for their support during my research work. I acknowledge all staff, research scholars, juniors and seniors of CSE Department, NIT Rourkela, India for helping me during my research work. I am grateful to NIT Rourkela, India for providing me adequate infrastructure to carry out the present investigations.

I take this opportunity to express my regards and obligation to my family members whose support and encouragement I can never forget in my life.

NIT Rourkela

Abinash Tripathy
Roll Number: 512cs1015

Abstract

Before buying a product, people usually go to various shops in the market, query about the product, cost, and warranty, and then finally buy the product based on the opinions they received on cost and quality of service. This process is time consuming and the chances of being cheated by the seller are more as there is nobody to guide as to where the buyer can get authentic product and with proper cost. But now-a-days a good number of persons depend upon the on-line market for buying their required products. This is because the information about the products is available from multiple sources; thus it is comparatively cheap and also has the facility of home delivery. Again, before going through the process of placing order for any product, customers very often refer to the comments or reviews of the present users of the product, which help them take decision about the quality of the product as well as the service provided by the seller. Similar to placing order for products, it is observed that there are quite a few specialists in the field of movies, who go though the movie and then finally give a comment about the quality of the movie, i.e., to watch the movie or not or in five-star rating. These reviews are mainly in the text format and sometimes tough to understand. Thus, these reports need to be processed appropriately to obtain some meaningful information. Classification of these reviews is one of the approaches to extract knowledge about the reviews. In this thesis, different machine learning techniques are used to classify the reviews. Simulation and experiments are carried out to evaluate the performance of the proposed classification methods.

It is observed that a good number of researchers have often considered two different review datasets for sentiment classification namely aclIMDb and Polarity dataset. The IMDb dataset is divided into training and testing data. Thus, training data are used for training the machine learning algorithms and testing data are used to test the data based on the training information. On the other hand, polarity dataset does not have separate data for training and testing. Thus, k-fold cross validation technique is used to classify the reviews. Four different machine learning techniques (MLTs) viz., Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Linear Discriminant Analysis (LDA) are used for the classification of these movie reviews. Different performance evaluation parameters are used to evaluate the performance of the machine learning techniques. It is observed that among the above four machine learning algorithms, RF technique yields the classification result, with more accuracy.

Secondly, n-gram based classification of reviews are carried out on the aclIMDb dataset.

The different n-gram techniques used are unigram, bigram, trigram, unigram+bigram, bigram + trigram, unigram + bigram + trigram. Four different machine learning techniques such as Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD) techniques are used to classify the movie reviews based on the n-gram technique as mentioned earlier. Different performance evaluation parameters are used to evaluate the performance of these machine learning techniques. The SVM technique with unigram + bigram approach has shown more accurate result among all other approaches.

Thirdly, SVM-based feature selection method is used to select best features from the set of all features. These selected features are then considered as input to Artificial Neural Network (ANN) to classify the reviews data. In this case, two different review datasets i.e., IMDb and Polarity dataset are considered for classification. In this method, each word of these reviews is considered as a feature, and the sentiment value of each word is calculated. The feature selection is carried out based on the sentiment values of the phrase. The words having higher sentiment values are selected. These words then act as an input to ANN on the basis of which the movie reviews are classified.

Finally, Genetic Algorithm (GA) is used to represent the movie reviews in the form of chromosomes. Different operations of GA are carried out to obtain the final classification result. Along with this, the GA is also used as feature selection to select the best features from the set of all features which eventually are given as input to ANN to obtain the final classification result. Different performance evaluation parameters are used to evaluate the performance of GA and hybrid of GA with ANN.

Sentiment analysis often deals with study of reviews, comments about any product, which are mostly textual in nature and need proper processing to obtained any meaningful information. In this thesis, different approaches have been proposed to classify the reviews into distinct polarity groups, i.e., positive and negative. Different MLTs are used in this thesis to perform the task of classification and performance of each technique is evaluated by using different parameters, viz., precision, recall, f-measure and accuracy. The results obtained by the proposed approaches are found to be better than the results as reported by other authors in literature using same dataset and approaches.

***Keywords*: *Sentiment Classification*; *IMDb Dataset*; *Polarity Dataset*; *Machine Learning Algorithms*; *Performance Evaluation Parameters*.**

Contents

List of Figures

List of Tables

List of Abbreviations

MLTs	Machine Learning Techniques
SA	Sentiment Analysis
EPF	Employee Provident Fund
PLSA	Probabilistic latent semantic analysis
IMDb	Internet Movie Review Database
NB	Naive Bayes
SVM	Support Vector Machine
RF	Random Forest
LDA	Linear Discriminant Analysis
ME	Maximum Entropy
SGD	Stochastic Gradient Descent
ANN	Artificial Neural Network
GA	Genetic Algorithm
NeuroGA	Neuro Genetic Algorithm
KNN	K Nearest Neighbor
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
POS	Part of Speech
NLP	Natural Language Processing
UCI	UC Irvine Machine Learning Repository
TF-IDF	Term Frequency Inverse document Frquency
CV	Countvectorizer
NPV	Negative Predictive Value
TNR	True Negative Rate
BOW	Bag of Words
CBOW	Continuous Bag of Word
IG	Information Gain
BGV	Best Gene Vector

Chapter 1

Introduction

In the recent years, with the increase in obtaining reviews, comments or sentiments from a number of on-line marketing and social networking sites, it is observed that very often customers or users express their idea, experience about any product or any news. Thus, these reviews become a source of information gathering for the new users or producers or sales managers. They get an opportunity to obtain detail information about the quality of the product, which helps them to take right decision to buy or produce or sell the product or not. Similarly for the case of movies, people give their comment about the quality of the movie. The issues related to these reviews are that they are mostly in the text format and hence, they need proper processing to obtain any meaningful information. Sentiment analysis performs this task by processing these review and classifies or clusters them depending upon the requirement of the users [1].

The rest of the chapter is organized as follows:
Section 1.2 provides an introduction to the sentiment analysis approach. Section 1.2 discusses about the sentiment analysis, its types, and its applications. Section 1.3 presents some challenges in the field of sentiment analysis. Section 1.4 highlights a brief information about different machine learning techniques (MLTs). Section 1.5 presents the motivation of the thesis work. Section 1.6 indicates the objectives of the work. Section 1.7 discusses about the thesis contributions. Section 1.7 presents the summary of the chapter.

1.1 Introduction

The comments or reviews or sentiments are mostly available in the social media and different on-line sites to help users gain knowledge about the item or topic. Thus, these reviews perform an appropriate role in decision making. According to two surveys of more than 2000 American adults, it is found out that [2, 3]:

- 73% to 87% of the frequent travelers, who go through on-line reviews of hotels, restaurants, and other services, report that these reviews have a significant influence on their purchase.
- 32% of people have provided a rating on any product or service, and 30% have posted on-line comment or review regarding any product.

- 81% of internet users have performed on-line research on any product at least once.
- Consumers willing to pay 20% to 99% more for a five star rated item to a four star rated item.

Thus, people not only prefer to write a comment about any topic but also like to go through the reviews while buying any product or using any service. But, these reviews need to be processed to obtain any commonly acceptable meaningful information about the topic. Hence, the role of sentiment analysis becomes important as it collects these reviews, processes them and finally helps the people to take any decision related to the topic.

1.2 Sentiment Analysis

Sentiment analysis (SA) analyzes people's opinions or reviews towards any product, organization, and their attributes, to generate a meaningful information [1]. These reviews are mainly in the text format and mostly unstructured in nature. Thus, these reviews need to be processed appropriately to obtain any meaningful information. Sentiment analysis is also known as opinion mining, opinion analysis, Subjectivity analysis, and emotional analysis. The term SA was first used by Nausaka and Yi [4] and the term opinion mining was first employed by Dave *et al.* [5]. But earlier to this, Elkan has a patent on text classification includes sentiment, humor and other concepts such as class labels [6]. The word sentiment denotes the underlying positive or negative feeling implied by a review. Thus, SA focuses on studies that indicate positive or negative sentiments.

1.2.1 Different Levels of Sentiment Analysis

Sentiment analysis is mostly carried out in different levels of granularity, which can be described as follows [7]:

- Document level sentiment analysis: The whole document is considered as a single unit. While processing the reviews, the analysis from entire document is either found as of positive or negative polarity [8, 9]. This level of analysis assumes that the whole document expresses the opinion on a single entity, but it is not useful for documents which access multiple objects. For such type of cases more fine level of granularity analysis needs to be carried out.
- Sentence level sentiment analysis: Each sentence is analyzed to check its polarity, i.e., either positive, negative. Neutral opinion is equivalent to no opinion. This analysis is comparable to that of subjectivity classification, which intends to separate the sentences based on precise information from the sentences and expresses them as subjective views [10].

- Aspect level sentiment analysis: Both document and sentence level SA observe the reviews of like or dislike categories. They do not represent the target of the reviews. To obtain this level of SA, a fine granular level of analysis is needed. This level of analysis is previously known as feature level SA [11, 12]. The aspect level analysis directly looks at the opinion and its target. The goal of this level of analysis is to discover sentiment on entities and their aspects.

- Comparative sentiment analysis: Sometimes people do not provide any direct review about any product; rather they give a comparison of the product with any other product. Identifying and extracting preferred entries about these reviews are considered as comparative SA. Jindal and Liu have provided an evaluation mechanism to handle comparative analysis of the sentiment [13]. They first identified comparative sentences present in the reviews and then tried to represent them in a relationship as follows: (<relation word>, <features>, <entity1>, <entity2>).
  The representation of an sentence can be explained considering an example as:
  Sentence: Mi's camera is better than that of Nokia.
  Representation: (<better>,<camera>,<Mi>,<Nokia>)

- Sentiment Lexicon acquisition: As discussed in comparative SA, it is found out that sentiment lexicon is the valuable resource for SA. According to Feldman, there exists three different ways to obtain the sentiment lexicons [7]. The different sentiment lexicon techniques are as follows:

  1. Manual approach: In this type of plan, people select the sentiment lexicon manually. This method is not feasible as for each domain a different set of lexicons need to be found out and for those, different domain experts are needed.
  2. Dictionary based approach: In this type of approach, a set of words associated with sentiments are initially considered and then, the set is expanded using the help of wordnet [14]. The final set of sentiment lexicons is identified having selected set of words associated with sentiments along with its synonyms and antonyms.
  3. Corpus-based approach: In this type of approach, a large set of texts related to the topic, called as corpus is considered. Like the dictionary based approach, a set of sentiment lexicons are initially found out; then the set is expanded using the corpus.

1.2.2 Various Application of Sentiment Analysis

Few of SA applications are discussed below [15]:

- Decision Making: Long before, when the opinions of users were not available on-line or publicly available, new users used to search for the users, who were using the

product, for the query related to the product. This is a cumbersome task to find out the old users and again to get a comment from them. But nowadays the present day users of the product share their views about the product, on-line through the social media or the on-line purchasing sites which help the new users to take a decision on using or buying the product.

- Reshaping Business and Control Public Sentiment: Different blogs or forum posts are maintained by both companies and government organizations to study view or suggestion about their existing product and also for the improvement in future products. Even the government agencies also consider users' feeling towards the new rules set by them. For example: Recently the Finance Minister of the Government of India during his presentation of the budget for the financial year 2016-17, inform that they plan to impose a tax on Employee Provident Fund (EPF). This proposal was very much criticized by the general public in social media and different forums which finally leads to the withdrawal of the proposal. Thus, the analysis of sentiment of people helps the organization and government agencies to change or modify their proposed rules for the betterment of system.

- Movie Success and Box-office Revenue: Along with the real-life application, many application oriented research work have been carried out in the field of SA. A good number of authors have proposed the articles in the area of movie reviews and box office collection. Mishne and Glance have indicated that positive feeling is a better predictor of the movie success [16], while Sadikov *et al.*, have made the same prediction using sentiment and other features [17]. Liu *et al.*, have proposed an approach for feeling model for predicting the box-office collection [18]. Their approach consists of two steps. The first step, constructs a model based on the probabilistic latent semantic analysis (PLSA) using only words associated with sentiment in the movie review dataset. The second phase, creates an autoregressive model using both revenues and opinion topics of last few days to predict the future revenue. Asur and Huberman have also performed the same prediction, but by tweet volume and tweet sentiment [19].

- Electoral Predictions: A good number of authors have used the concept of evaluation of the opinions of public, for predicting the electoral result. O'Connor *et al.*, have computed sentiment score by counting the words having positive and negative polarity, correlating those result and finding out a better accuracy [20]. Bermingham and Smeaton have used tweet volume for prediction. They have considered the positive and negative tweets as independent variables and polling result as a dependent variable to train a linear regression model to predict the election result [21]. On the other hand, Diakopoulos and Shamma [22] and Sang and Bos [23] have proposed manual annotation of sentiments of tweets for prediction of the election result.

- Stock Market Prediction: Another popular application area of SA is a stock market prediction. Das and Chen have considered the message board posts and then have selected opinions from those posts to classify them into three different classes such as bullish (optimistic), bearish (pessimistic), or neutral [24]. They have collected sentiment about all the stocks, then combined them and finally predicted the Morgan Stanley High-Tech Index. Zhang *et al.*, have obtained positive and negative moods on Twitter and then they have used them for prediction of stock market indices for Dow Jones, S&P 500, and NASDAQ [25]. Bar-Haim *et al.*, have identified expert investors based on past prediction of bullish and bearish stocks[26]. They have considered the opinions of these experts as features and based on these features they have predicted the stock market indices Si *et al.*, have combined topics based on sentiment time series and index time series to predict the S&P 100 index's daily movement using vector of auto regression [27].

1.3 Challenges in Sentiment Analysis

Sentiment analysis is mainly concerned with processing the reviews, comment on different people and processing them to obtain any meaningful information from it [15]. Different factors affect the process of SA, and need to be handled properly to get the final classification or clustering report. Few of these challenges are discussed below [28]:

- Co-reference Resolution: This problem is mainly referred to finding out ''what does a pronoun or proverb indicate ?'' For example, in sentence ''After watching the movie, we left for food; it was good.'' What does the word ''it'' refers to; whether the movie or food? Thus, when the analysis about the movie is being carried out, whether the sentence relates to movie or food ? This is a concern for the analyst. This type of issue mainly occurs in the case of aspect-oriented SA.

- Association with a period: The time of opinion or review collection is an important issue in the case of SA. The same user or group of users might give a positive response for a product at a given time, and there might be a case where they might give a negative response. Thus, it is a challenge for the sentiment analyzer at some other instance of time. This type of issue mainly occurs in comparative SA.

- Sarcasm Handling: The use of words which mean opposite to what they inform are mostly known as sarcasm words. For example, the sentence ''What a good batsman he is, he scores zero in every other innings.'' In this case, the positive word ''good'' has a negative sense of meaning. These sentences are tough to find out and thus, they affect the analysis of the sentiment.

- Domain Dependency: In SA, the words are mainly used as a feature for analysis. But, the meaning of the words is not fixed through out. There are few words whose

meanings change from domain to domain. Apart from that, there exists words which have opposite meaning in different situations known as contronym. Thus, it is a challenge to know the context for which the word is being used, as it affects the analysis of the text and finally the result.

- Negations: The negative words present in a text can totally change the meaning of the sentence in which it is present. Thus, while analyzing the reviews, these words need to be taken care of. For example, The sentences ‘‘This is a good book.’’ and ‘‘This is not a good book.’’ have opposite meaning, but when the analysis is carried out using the single word at a time, the result may be different. To handle this type of situations, n-gram analysis preferred.

- Spam Detection: Sentiment analysis is concerned with the study of reviews. But, till date very little qualitative analysis has been made for checking as to whether the reviews are fake, or any valid person has given the review. Many people without any knowledge of the product or the service of the company provide a positive review or negative review about the service. This is very much difficult to check as to which review is a fake one and which is not; that eventually plays a vital role in SA.

1.4 Machine Learning Techniques

A computing machine can only understand the general representation of text, if it is represented properly. Thus, the texts of the reviews need to be converted into a proper format to instruct a machine. Again, the machine understands or learns a specific set of data called training data and based on the learning of training data, predicts the other set of data, i.e., the untrained or testing data. Machine learning techniques (MLTs) help in learning as well as predicting. The various types of MLTs can be explained as follows:

- Supervised MLTs: This is the most commonly used MLT. In this type of learning, both the training and testing data are labeled, i.e., each text file of the dataset has a polarity value assigned to them viz., positive or negative or neutral. The training dataset is used by the system for training, and based on this information, the testing data is labeled [29]. As the testing dataset already has a label, both the labels are compared to obtain the final accuracy of the system.

- Unsupervised MLT: This type of MLT does not have a labeled dataset. Thus, while analysis of these reviews, clustering approach is considered, which makes a group of similar types of the elements into a cluster [30]. Various different evaluation parameters are considered to check the performance of these techniques.

- Semi-supervised MLT: In this type of approach, a small size of label dataset is present, where the size of the unlabeled dataset is large [31]. Thus, using the small size labeled

dataset, this approach makes an attempt to label the whole dataset. The small labeled dataset is trained and based on these values a small size of the unlabeled dataset is predicted. These predicted data are added to the already labeled dataset until the total data is labeled.

1.5 Motivation

The motivation for this research work can be explained as follows

- Since sentiment analysis is concerned with the study of reviews, opinions on any topic and providing meaningful information, selecting a proper authentic set of reviews for processing is a challenging job. Thus, the reviews considered for analysis, which is mainly used by different authors for analysis and classification
- The reviews or comments provided by the people are mainly in the text format which is sometimes tough to understand and process. Thus, a proper preprocessing mechanism needs to be adopted to remove unwanted, confusing information for the data sets. Hence, different mechanisms like stop word, numerical and special character removal, which do not play any active role in sentiment analysis of the texts and along with this all text are converted into either lower or upper case, to maintain uniformity during the analysis of the reviews.
- Different MLTs help to classify or cluster the reviews. These reviews need to be represented in the form of numerical values, which are considered by MLTs for input. The conversion of text reviews into a numerical values is a challenge, and it needs to be processed properly for finding a conclusion. Hence, mainly two different techniques like Countvectorizer and TF-IDF are used, which converts the texts into numerical vector based on the occurrence and, both occurrence and their number of occurrence respectively.
- In case of SA, each word is considered as a feature. But as discussed in Section 1.3, the negative comments in reviews play an important role in SA. So, consideration of a single word (unigram) does not provide good result always. In such situations, n-gram approach is needed, i.e., the collection of two or three words as a single unit, which is also known as bigram or trigram respectively.
- Sentiment analysis is mainly concerned with the study of reviews or opinions. These reviews are in text formats. Each word of these reviews can be considered as a feature for analysis. It is observed that sometimes the collection of all words becomes vast and it may contain words which may not affect the sentiment of the reviews. Thus, a feature selection mechanism needs to be adopted to select the best features out of all the features, which affect the sentiments of text.

- • In this thesis, different machine learning techniques are used in different chapters. This is done as while analyzing the literatures in that area, the techniques used by the authors, who have carried out the analysis are first preferred and along with this another techniques are used, which do not use the same approach for analysis.

In this thesis, an attempt has been made to analyze the sentiment of movie reviews using different classification methodologies.

1.6 Objectives

In this thesis, some of the challenges related to SA are considered with a focus on classification of reviews or opinions in a best possible way. The main objective of this research work can be outlined as:

i. To consider an authentic review dataset or opinion set for analysis and check whether the approach is valid for all similar kinds of datasets or not.

ii. To pre-process the dataset before the analysis starts by removing unwanted words.

iii. To convert the text reviews into a matrix of numerical values that act as input to MLTs for sentiment analysis.

iv. To classify the review by not only a single word but also collection of two words (bigram) or three words (trigram) as a single unit to obtain the best possible result after classification result.

v. To use proper feature selection mechanism to select the best features from the set of all features, which have a significant effect on the sentiment of the reviews.

1.7 Thesis Contribution

The contribution of this thesis can be explained as follows:

Chapter 3 proposes analysis of movie reviews in the form of classification, using different MLTs. Two different datasets i.e., Internet Movie Database (IMDb) [32] and Polarity dataset [33] are used for classification. These two datasets are considered for analysis as most of the authors have chosen and analyzed these dataset for classification purpose. Four different MLTs viz., Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Linear Discriminant Analysis (LDA) are used to classify the movie reviews. These methods are considered for analysis as NB and SVM are used by most of the authors, while NB uses probabilistic Bayesian method and SVM uses kernel based approach for analysis. In this chapter, RF and LDA are also used for analysis as RF uses ensemble method and LDA uses discriminant analysis approach for classification. All these methods

are preferred to test whether the propose approach work in all environment and found out that the approach work fine in all environment. The IMDb dataset has separate data for testing and training. Thus, a training dataset is used for training and based on that, the testing on the testing dataset is carried out. While polarity dataset does not have separate data for training and testing, thus, k-fold cross validation technique is used for classification. The RF technique shows the best result on both datasets among the four different MLTs.

Chapter 4 proposes a classification of movie reviews using n-gram machine learning techniques. The IMDb movie review dataset is considered for classification unlike the chapter 3 where two different datasets are considered as the polarity dataset classification work on the principle of k-fold cross validation technique and where n-gram technique is used there, it makes the processing more complex. The IMDb movie review dataset is considered for classification. Different n-gram techniques used are unigram, bigram, trigram, unigram + bigram, bigram + trigram, unigram + bigram + trigram. Four different MLTs are used for classification viz., NB, Maximum Entropy (ME), SVM, Stochastic Gradient Descent (SGD). These MLTs are preferred, as different authors have used NB, SVM and ME frequently for analysis. Again, SGD is used as it works on the principle of gradient descent for analysis and it helps to check whether the proposed approach work in all approaches. Different performance evaluation parameters are used to evaluate the performance of the classifier and SVM with unigram + bigram approach shows the best result among all other approaches.

Chapter 5 proposes sentiment classification using Genetic Algorithm (GA) and NeuroGenetic algorithm (NeuroGA), i.e., the hybrid form of ANN and GA on polarity movie review dataset. . A hybrid approach is preferred in this chapter in order to avoid the bias of any particular technique on the dataset. In this type of processing, each test review is represented as a chromosome. GA takes this chromosome as input, and then using different GA operations, it helps to classify the text reviews into different polarity groups. Apart from the hybrid approach, the GA used to classify the reviews using its different operators and shows an accuracy of 93%. Again GA is used to select the best features from the set of all features. These selected features are then given as input to ANN, which classifies the reviews into different polarity groups. During the process of ANN classification, the hidden nodes are kept on changing to find out the best possible result. The performance of the classifier is evaluated using different parameters.

Chapter 6 proposes sentiment classification using SVM and Artificial Neural Network (ANN) on both IMDb and polarity dataset. SVM is used for feature selection process, which calculates the sentiment value of each word and then finally considers a threshold value of sentiment to select the best features from the set of all features. In this chapter, the threshold value is set to be mod(0.009) of the obtained sentiment values of each word. These selected features are then given as input to ANN and based on these features; ANN classifies the testing reviews. The inputs being the best-selected features; the output is obtained for two

different classes while the hidden nodes are kept on changing to find out the best possible classification result. Different performance evaluation parameters are used to evaluate the performance of the proposed approach.

Chapter 7 discusses sentiment clustering using unsupervised machine learning techniques. A chapter on unsupervised approach is added to the thesis as collecting the labeled dataset is a difficult task and thus, the approach on unlabeled twitter data is carried out. The reviews for analysis are collected from Twitter using Twitter API. Then, these reviews are clustered in two different clusters i.e., positive and negative cluster using different unsupervised clustering algorithms namely K means, mini batch K means, Affinity propagation, and DBSCAN. Four different clustering algorithms are preferred in this chapter as all fours works on different principle such as K means works on centroid selection mechanism, mini batch K means works on small dataset, Affinity propagation works on the principle of similarity between the inputs and DBSCAN works on the principle of density of the input points. The performance of these techniques are evaluated by using different performance evaluation parameters like Homogeneity, Completeness, V-measure, Adjust Rand Index, and Silhouette Coefficient.

1.8 Thesis Organization

This thesis is organized into eight different chapters including the introduction section. Each of the chapters is discussed below briefly.

Chapter 2: Literature Survey

This chapter focuses on the state-of-art of various sentiment classification methods. The first section provides a survey of sentiment classification methods. The second section provides a survey of sentiment classification methods using n-gram techniques. The third section provides a study on the use of hybrid MLTs for classification. The fourth section discusses a study on the use of different feature selection mechanism for sentiment classification.

Chapter 3: Classification of Sentiment of Reviews using Supervised Machine Learning Techniques

This chapter proposes sentiment classification technique using four MLTs on two different datasets, i.e., IMDb [32] and Polarity [33]. With the unavailability of separate dataset for training and testing, 10 fold cross validation technique is used for classification in Polarity dataset, while as the dataset is separated into training and testing in IMDb dataset, the training data used for training and based on that information the testing dataset is classified. The performance evaluation parameters are used to check the performance of different machine learning techniques.

Chapter 4: Classification of Sentiment Reviews using N-gram Machine Learning Approach

This chapter proposes sentiment classification using n-gram MLTs. Four MLTs are used along with n-gram techniques like unigram, bigram, trigram, unigram + bigram, bigram + trigram, unigram + bigram + trigram to classify movie reviews of the IMDb dataset. To evaluate the performance of MLTs different performance evaluation parameters are used and SVM with unigram + bigram approach shows the best result among all other approaches.

Chapter 5: Document level Sentiment Analysis using Genetic Algorithm and Neuro-Genetic Algorithm

This chapter proposes a classification of sentiment reviews using GA and NeuroGA methods. For GA analysis, the text reviews are represented as in the form of chromosomes. The different operations of GA are performed on this chromosomes and finally classification result is obtained. Again GA is used for feature selection. The selected features are then given input to ANN, by which is classify the testing data. The performance of this approach can be evaluated using different performance evaluation parameters.

Chapter 6:Document level Sentiment Classification using Feature Selection Technique

This chapter proposes a hybridization of SVM and ANN techniques on two different datasets, i.e., IMDb and Polarity. The sentiment value of each word / features is calculated using SVM. Then, a threshold sentiment value is considered and the features that have higher sentiment values are only considered. These selected features are then given input to ANN, by which the testing data are being tested. Different evaluation parameters are used to check the performance of the proposed approach.

Chapter 7: Sentiment clustering using Unsupervised machine learning techniques

This chapter proposes clustering of Tweeter reviews collected using Twitter API. Four different machine learning techniques namely K means, Mini batch K means, Affinity propagation, and DBSCAN used to cluster the tweets collected Tweeter. Different performance evaluation parameters, i.e., homogeneity, completeness, V-measure, Adjusted Rand Index, Silhouette Coefficient are used to evaluate the performance of these techniques.

Chapter 8: Conclusion

This chapter presents a conclusive remark on the thesis based on the work done. The scope of future work is also discussed at the end.

Chapter 2

Literature Survey

This chapter discusses the research work performed by different researchers in the field of sentiment analysis. It describes different classification techniques to classify the reviews into different polarity groups, i.e., negative and positive polarity. It also focuses on the use of different hybrid MLTs for classification and also features selection techniques to select the best features from the set of significant feature and based on these selected features perform classification.

The rest of the chapter is organized as follows:

Section 2.1 provides a introduction to this chapter. Section 2.2 is concerned with document level sentiment classification using different MLTs on various datasets. Section 2.3 discusses classification of movie reviews using n-gram techniques. Section 2.4 presents the classification techniques using the different hybrid approach of MLTs. Section 2.5 is concerned with various feature selection methods employed in the area of sentiment classification. Section 2.6 discusses about the unsupervised approach for sentiment analysis i.e., clustering of the reviews. Section 2.7 is concerned with sentiment analysis using semi supervised approach. Finally, Section 2.8 provides the summary of the chapter.

2.1 Introduction

In this chapter, the different sentiment classification methods based on document level are discussed. As mentioned in section 1.2.1, document-level sentiment analysis considers the whole document as a single unit and then tries to classify it into either positive or negative polarity. Different MLTs are used for classification, but before the use of the MLTs, different steps for preprocessing data are carried out on the text reviews. Different approaches are adopted by authors to increase the accuracy of the system such as n-gram methods, feature selection, and use of hybrid MLTs.

2.2 Document Level Sentiment Classification

The document-level sentiment analysis considers the whole document as a single unit to analyze its polarity, i.e., either of positive or negative, or neutral. Important and related

articles on this topic, are discussed in this section.

- Pang and Lee have used polarity dataset for sentiment classification [33]. They have categorized the reviews into subjective and objective portions. They have considered only the subjective portion, while classification as objective portion does not contain any information about the sentiment. They have adopted the minimum-cut formulation in graph approach to obtain the subjective portion from the total text for review. They have used SVM and NB classifier for classification of reviews along with minimum cut formulation.

- Salvetti *et al.*, have discussed on overall opinion polarity (OvOp) concept using machine learning algorithms for classification of reviews [34]. They have used Naive Bayes and Markov Model techniques for classification. In this paper, the hypernyms have been provided by wordnet and Part Of Speech (POS) tag acts as the lexical filter for classification. They have suggested that the result obtained by wordnet filter is less accurate in comparison with that of POS filter.

- Beineke *et al.*, have used Naive Bayes model for sentiment classification. They have extracted a pair of derived features which are linearly combinable to predict the sentiment [35]. To improve the accuracy level, they have added additional derived features to the model and used labeled data to estimate relative influence. Along with this, they have also used the concept of anchor words, i.e., the words with multiple meaning for analysis. They have considered five positive anchor words and five negative anchor words which after combination produce 25 possible pairs for analysis. They have followed the approach of Turney, which effectively generates a new corpus of label document from the existing document [9].

- Mullen and Collier have applied SVM algorithm for sentiment analysis where values are assigned to few selected words and then combined them to form a model for classification [36]. Along with this, different classes of features having a closeness to the topic are assigned with the favorable values, which help in classification. The authors have presented a comparison of their proposed approach with data, having topic annotation and hand annotation. Their proposed method has shown better result compared to that of topic annotation whereas the results need further improvement while comparing with hand annotated data.

- Zhang *et al.* have proposed a rule based approach for classification of the reviews [37]. Their approach consists of two phases, i.e., sentence sentiment analysis and document sentiment aggregation. They decompose the document into its constituent sentences and find out polarity of each sentence. Then, polarity score of all sentences are combined to compute the overall polarity of the document. They have

considered Euthanasia dataset consisting of 851 Chinese articles and AmazonCN dataset consisting of 458,522 reviews from six different categories i.e., books, music, movie, electrical appliance, digital product, and camera. They have used SVM, NB and Decision tree techniques to classify the reviews.

- Yessenalina *et al.* have proposed a joint two-level approach for document level sentiment classification [38]. Their approach extracts the subjective sentences from the text and based on these sentences, the document is classified. Their training method considers each sentence as a hidden variable and jointly learns to predict the document label which controls the propagation of incorrect sentence labels. In order to optimize the document level accuracy, their model solves the sentence extraction subtask only up to the extent required for accurately classify the document sentiment. They have evaluated the movie reviews dataset [33] and U.S. Congressional floor debate dataset [39] for classification using SVM machine learning technique.

- Tu *et al.* have used sequence and convolution kernels using different types of structures for document level sentiment classification [40]. They use both sequence and convolution kernels for analysis. For sequence kernels, they have used a sequence of lexical words (SW), POS tags (SP) and combination of sequence of words and POS (SWP). For dependency kernel, they have used word (DW), POS (DP), and combined word and POS settings (DWP), and similarly for simple sequence kernels (SW, SP and SWP). They used vector kernel (VK) in a bag-of-words as baseline. Their approach of VK + DW has shown the best result among all the proposed result. They have used polarity [33] dataset for analysis.

- Bollegala and Carroll have proposed cross domain sentiment classification problem, which focuses on training the classifier from one or more domains and applying the trained classifier on another domain [41]. They have created a sentiment sensitive distributional thesaurus using labeled data. They have obtained sentiment sensitivity in the thesaurus by adding document label sentiment labels in the context vector, which is used to measure the distributional similarity between words. In order to reduce the mismatch between the features of different domains, they have appended additional related features to the feature vectors and their approach has shown comparably better result in the field of information retrieval and document classification. They have collected reviews from different domains, i.e., books from amazon.com, hotels from tripadvisor.com, movies from imdb.com, automobile from caranddriver.com and restaurants form yelp.com for classification. They have used L1 regularized logistics regression for classification of the reviews collected from different sources.

- Moraes *et al.* have compared the SVM and NB approach with ANN for document level sentiment classification [42]. They have used information gain (IG) approach to select

the best term from the reviews. The IG is mainly based on the number of occurrences of the term in the reviews. The higher IG score is given to the most frequently used words in the text. These words are then given input to the MLTs for classification. Among the three MLTs, ANN shows the best result. They have used Polarity dataset [33] and the reviews collected from Amazon based on the product like GPS, books and camera for sentiment classification.

- Tang has encoded the relationship between the sentence and the document while sentiment classification [43]. His approach is based on the principle of constitutionality, which indicated that, the meaning of a document can be derived from the meaning of its constituents and the rule used to combine them. He has proposed a model which learns the sentence representation using convolutional Neural Network. Then, semantics of the sentence and relationship between them are encoded in document representation which is finally considered for classification. He has considered four large scale review datasets from IMDb and Yelp challenge dataset for classification.

- Zhang *et al.* have proposed the classification of Chinese comments based on word2vec and SVM^{perf} [44]. Their approach is based on two parts. In the first part, they have used word2vec tool to cluster similar features in order to capture the semantic features in selected domain. Then in the second part, the lexicon based and POS based feature selection approaches are adopted to generate the training data. Word2vec tool adopts continuous bag-of-words (CBOW) model and continuous skip-gram model to learn the vector representation of words [45]. SVM^{perf} is an implementation of SVM for multi-variate performance measures, which follows an alternative structural formulation of SVM optimization problem for binary classification [46].

- Liu and Chen have proposed different multi-label classification on sentiment classification [47]. They have used eleven multilevel classification methods, with two micro-blog datasets and eight different evaluation matrices for analysis. Apart from that, they have also used three different sentiment dictionaries for multi-level classification. According to the authors, the multi-label classification process performs the task mainly in two phases, such as, problem transformation and algorithm adaptation [48]. In problem transformation phase, the problem is transformed into multiple single-label problems. During the training phase, the system learns from these transformed single label data, and in the testing phase, the classifier after learning process, makes a prediction at a single label, and then translates it to multiple labels. In algorithm adaption, the data is transformed as per the requirement of the algorithm.

- Luo *et al.*, have proposed an approach to convert the text data into low dimension emotional space (ESM) [49]. They have annotated small size words, which have

definite and clear meaning. They have also used Ekman Paul's research to classify the words into six basic categories such as anger, fear, disgust, sadness, happiness and surprise [50]. They again have considered two different approaches for assigning weight to words by emotional tags. The total weight of all emotional tags are calculated and based on these values; the messages are classified into different groups. Although their approach yields reasonably a good result for the stock message board, the authors claim that it can be applied to any other dataset or domain.

- Niu *et al.* have introduced multi view sentiment analysis dataset including as set of image-text pair with manual annotation collected from Twitter [51]. They have categorized the sentiment analysis into two categories, i.e., lexicon based approach and statistical learning approach. In lexicon based approach, a set of sentiment score is assigned to pre-defined words or phrases. The sentiment score of the text is the aggregation of sentiment score of each words in the text. They have also used some natural language processing (NLP) techniques to solve the issues related to syntax, negation and irony. In statistical learning approach, they have used SVM for classification of the reviews.

- Xia *et al.* have proposed a three-stage model for multilevel classification process where the stages are Polarity Shift Detection, Elimination and Ensemble (PSDEE) [52]. Firstly, they have employed a rule-based method to detect some polarity, and a statistical method to detect some implicit polarity shifts. Secondly, they propose a novel polarity shift elimination algorithm to eliminate polarity shifts in negations which makes the BOW representation more feasible. Finally, they separate the training and test data into four component subsets, i.e., negation subset, contrast subset, sentiment-inconsistency set as well as polarity unshifted subset, and train the base classifiers based on each of the component subset. They evaluate their model by conducting experiments on four sentiment datasets, three kinds of classification algorithms and two types of features. They have used a multi domain dataset by Blitzer *et al.* which comprises of four domains, i.e., Book, DVD, Electronics and Kitchen [53] for classification. They have used linear SVM, logistic regression and Naive Bayes with unigram and combination of unigram and bigram for classification.

Table 2.1 provides a comparative study of different approaches adopted by various authors as reported in literature on the topic of document-level sentiment classification.

Table 2.1: Comparison of Document level Sentiment Classification

Author	**Approach Considered**	**Algorithm Used**	**Obtained result (Accuracy %)**	**Dataset used**
Pang and Lee [33]	Considered minimum-cut formulation method on subjective document	Naive Bayes (NB) and Support Vector Machine (SVM)	NB: 81.5, SVMs: 65.9	Polarity Dataset
Salvetti *et al.* [34]	Accessed overall opinion polarity(OvOp) concept using machine learning algorithms	Naive Bayes (NB) and Markov Model (MM)	NB: 79.5, MM: 80.51	Internet Movie Database (IMDb)
Beineke *et al.* [35]	Considered linearly combinable paired feature are used to predict the sentiment	Naive Bayes	NB: 65.9	Internet Movie Database (IMDb)
Mullen and Collier [36]	Assined values to selected words then combined to form a model for classification	Support Vector Machine (SVM)	SVM: 86.0	Internet Movie Database (IMDb)
Zhang *et al.* [37]	Proposed rule based approach with sentence sentiment analysis and document sentiment aggregation phase for document level sentiment analysis	SVM, NB and Decision Tree	Euthanasia: SVM:83.88, NB:68, DT:76, AmazonCN: SVM:79.97, NB:73.53, DT:70.32	Euthanasia dataset and AmazonCN dataset.
Yessenalina *et al.* [38]	Extracted subjective sentences form the text and based on these sentences classification is carried out	SVM	Movie review: 92.67, US Congressional debate: 78.84	Movie review dataset and US Congressional floor debate dataset
Tu *et al.* [40]	Used sequence and convolution kernels using different types of structures for document level sentiment classification.	Vector Kernel(VK) with sequence and dependency kernel	VK + DW: 88.50	Polarity dataset
Bollegala and Carroll [41]	Proposed cross domain sentiment classification problem by adding additional features to feature vectors to reduce mismatch between different domains	L1 regularized Logistic Regression	overall: 80.91	Books from Amazon.com, Hotel from tripadvisor.com, Movies from imdb.com, automobiles from caranddriver.com and Restaurants from yelp.com
Moraes *et al.* [42]	Selected the terms using Information Gain ranking and then classify the reviews using NB, SVM and ANN	NB, SVM and ANN	NB:80.3, SVM : 84.1, ANN:86.5	Polarity dataset and Amazon reviews on GPS, Books and Camera.
Tang [54]	Their approach based on principle of constitutionality i.e., deriving meaning from the constituents and then use rule to combine them	CNN	CNN:86.58	Four large dataset from IMDb and Yelp challenge dataset.
Zhang *et al.* [44]	Used word2vec to capture similar features then classify reviews using SVM^{perf}	SVM^{perf}	Lexicon based: 89.95, POS based: 90.30	Chinese comments on clothing products
Liu and Chen [47]	Used multi-label classification using eleven state-of-art multi-label, two micro-blog datasets, and eight different evaluation matrices on three different sentiment dictionaries.	Eight different evaluation matrices	Average highest Precision: 75.5	Dalian University of Technology Sentiment Dictionary (DUTSD), National Taiwan University Sentiment Dictionary (NTUSD), Howset Dictionary (HD)
Luo *et al.* [49]	Used Ekman Paul's research approach to convert the text into low dimensional emotional space (ESM), then classify them using machine learning techniques [50]	Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT)	SVM: 78.31, NB: 63.28, DT: 79.21	Stock message text data(The Lion forum)
Niu *et al.* [51]	Used Lexicon based analysis to transform data into required format and then use statistical learning methods to classify the reviews	BOW feature with TF and TF-IDF approach	Text: 71.9, Visual Feature: 68.7, Multi-view:75.2	Manually annotated Twitter data
Xia *et al.* [52]	Used three-stage model, i.e., Polarity Shift Detection, Elimination and Ensemble (PSDEE), for document-level sentiment classification	SVM, logistic regression (LR), and Naive Bayes	SVM: 0.871, LR: 0.874, NB: 0.891	multi domain dataset by Blitzer *et al.* that comprises of four domain i.e., Book, DVD, Electronics and Kitchen [53]

2.3 Sentiment Classification using n-gram MLTs

During the process of sentiment classification, each word is considered as a feature and the word plays an important role in assigning the polarity to the document. When the analysis is carried out using a single word, it is called as *''Unigram''*. While for two consecutive words or three consecutive words, it is known as ''bigram'' and ''trigram'' respectively. For the value of n, i.e., number of consecutive words are more than three, four-gram or five-gram are used. A good number of authors have used this approach for classification of movie reviews. Few important and relevant articles are discussed below:

- Pang *et al.*, have considered the sentiment classification method based on categorization study, with positive and negative sentiments [8]. They have undertaken the experiment with three different machine learning algorithms, such as Naive Bayes, Support Vector Machine (SVM), and Maximum Entropy. The classification process is undertaken using the n-gram technique like unigram, bigram, and the combination of both unigram and bigram. They have also used bag-of-word (BOW) framework to implement the machine learning algorithms.

- Kešelj *et al.* have proposed a byte level n-gram author profile verification approach for author writing [55]. Their approach is observed to be independent of language, special character, and case. They have chosen the optimal set of n-gram to be included in the profile and then calculated the similarity between pairs. They have performed the experiment with three different languages dataset i.e., English, Greek, and Chinese to prove that their approach is language independent.

- Matsumoto *et al.*, have used the syntactic relationship among words as a basis of document-level sentiment analysis [56]. In their article, frequent word sub-sequence and dependency sub-trees are extracted from sentences, which act as features for SVM algorithm. They have extracted unigram, bigram, word subsequence and dependency subtree from each sentence in the dataset. They have used two different datasets for conducting the classification i.e., IMDb dataset [32] and Polarity dataset [33]. In the case of IMDb dataset, the training and testing data are provided separately but in Polarity dataset 10-fold cross validation technique is considered for classification as there is no separate data designated for testing or training.

- Bespalov *et al.*, have proposed the embedding of higher order n-gram phrases into low dimensional latent semantic space [57]. They have also used the deep neural network to build a discriminative framework which estimates the parameters of latent space and the classification function with a bias towards the classification task. They have also proposed both binary classification, i.e., the classification of reviews into two different classes and multi-score sentiment classification, i.e., predicting the sentiment

score within a range of values. Two benchmark datasets i.e., Amazon and TripAdvisor are used by them for classification.

- Ghiassi *et al.*, have proposed an approach for feature reduction using n-gram approach and performed statistical analysis to develop the Twitter-specific lexicon for sentiment analysis [58]. They have found out four different areas associated with Twitter sentiment analysis, i.e., data gathering, determining sentiment scale for data, feature engineering, and finally evaluation and classification of the Twitter message. They have developed a sentiment classification model using Twitter-specific lexicon and dynamic artificial neural network (DAN2) and compared the result of their proposed system with SVM. They have collected Twitter tweets and considered them for sentiment classification.

- Sidorov *et al.*, have proposed syntactic n-gram (sn-gram) technique [59]. In traditional n-gram technique, the neighboring words of the particular word are considered for analysis where as in sn-gram technique the neighboring word of a particular word is considered based on the syntactic relationship between words. They have discussed the use of sn-gram for authorship attribution. Along with sn-gram technique, they also used n-gram technique based on words, characters, and Part Of Speech (POS) tags. They have used three different machine learning techniques, i.e., SVM, NB, and tree classifier J48 for classification.

- Tang *et al.*, have proposed sentiment specific word embedding (SSWE) technique which considers sentiment related information in the continuous representation of the word, i.e., the n-gram technique [54]. They proposed three different versions of the neural network to incorporate the sentiment polarity of text in their loss function. They have proposed $SSME_u$, i.e., the unified model of SSWE, which considers not only the sentiment information but also the syntactic context of the word. $SSME_h$, i.e., the unsupervised approach, which does not consider the whole sentence, rather it considers a window of n-gram across the sentiment to predict sentiment. In $SSME_r$, the objective function is more relaxed in comparison with those of other two and does not require any probabilistic interpretation for sentiment classification.

- Agarwal *et al.*, have proposed a concept of the parser based on semantic information and the relationship between words [60]. They have used Maximum Redundancy and Maximum Relevance (mRMR) approach for feature selection. They have used unigram, bigram, bi-tagged and dependency parse tree for feature selection. The unigram considers one word, bigram considers two consecutive words, bi-tagged selectively extract based fixed pattern of POS, and dependency parse tree approach generates a parse tree of the sentence to select the best feature from the set of large feature. They have considered Cornell movie review dataset, which contains 2000

labeled movie reviews and Amazon product review dataset consisting review about books, DVD, and electronics for classification using SVM technique.

- Foroozan *et al.* have proposed an experiment based on n-gram approach used as feature extraction with different weighting methods [61]. They have considered unigram, bigram, and combination of unigram and bigram method along with linear kernel based SVM and RBF kernel based SVM for classification. In addition, they have used feature weighting method like binary, TF, TF-IDF with document frequency to represent each news files. They have collected 2308 news articles belonging to three different classes, i.e., positive (1160), negative (757) and neutral (391) from Google finance for classification.

- Aisopos *et al.* have proposed a supervised machine learning model to test experimental results with multilingual manually annotated post form Twitter [62]. They have used n-gram graph technique for classification of reviews. Their approach is both language and noise i.e., unwanted information agnostic which eventually improves the classification result. They have performed experiment on multilingual (Spanish, English, Portuguese, Dutch, and German) and multi topic tweets collected from twitter. They have collected 95608 tweets combining corpora available for research. They have also used Multinomial NB, SVM, Logistics Regression, C4.5 Tree, Multilayer perceptron and K Nearest Neighbor techniques for classification of the tweets.

Table 2.2 provides a comparative study of different approaches adopted by authors on sentiment classification using n-gram MLTS.

2.4 Sentiment Classification using Hybrid MLTs

In the Section 2.2 and Section 2.3 sentiment classification techniques using different MLTs have been discussed, where the authors have used more than one machine learning techniques, but they perform the task of classification independent of each other. But there are few articles where authors have not only used more than one machine learning techniques, but they are also related to each other. This type of combined approaches of MLTs to perform classification are known as hybrid approach. This section discusses on few of this categories of articles where hybrid MLTs approach is used. They are highlighted as below:

- Abbasi *et al.* have proposed a sentiment classification approach using entropy weighted genetic algorithm (EWGA) and SVM [63]. They have evaluated different feature sets consisting of syntactic and stylistic features. Stylistic features represent vocabulary richness measure, word-length distribution, and special character frequency. The EWGA approach uses information gain (IG) heuristic to assign weights to various sentiment attributes. These weights are then added to the initial

Table 2.2: Comparison of Sentiment Classification using n-gram MLTS

Author	Approach Considered	Algorithm Used	Obtained result (Accuracy %)	Dataset used
Pang *et al.* [8]	Classified the dataset using different machine learning algorithms and n-gram model	Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM)	Unigram: SVM (82.9), Bigram: ME (77.4), Unigram + Bigram : SVM (82.7)	Internet Movie Database (IMDb)
Kešelj *et al.* [8]	Proposed byte level n-gram verification approach for author writing.	N-gram approach	English: 83, Greek: 73, Chinese: 89	Dataset collected from three different languages i.e., English, Greek and Chinese.
Matsumoto *et al.* [56]	Used syntactic relationship among words as a basis of document level sentiment analysis	Support Vector Machine (SVM)	Unigram: 83.7, Bigram: 80.4, Unigram+Bigram : 84.6	Internet Movie Database (IMDb), Polarity dataset
Bespalov *et al.* [57]	Embedded high order n-gram and deep neural network to classify the reviews	SVM, Supervised latent n-gram analysis (SLNA) and Bag of words (BOW)	SLNA: 92.88, BOW + SVM: 92.63	Amazon, TripAdvisor
Ghiassi *et al.* [58]	Used Feature reduction using n-gram and statistical analysis on twitter specific lexicon for classification	SVM, Dynamic Artificial Neural Network (DAN2)	SVM: 94.6, DAN2: 95.1	Twitter (10,345,184 tweets)
Sidorov *et al.* [59]	Used syntactic n-gram technique, n-gram techniques for words, characters and POS tagged and different machine learning technique for classification.	SVM, NB, and J48	SVM : 93, NB : 90, J48 :86	Text downloaded from Project Gutenberg and books of native English speaking authors.
Tang *et al.* [54]	Used sentiment specific word embedding (SSWE) technique along with three different versions of Neural Network to classify reviews.	Three different versions of Neural network	$SSWE_u$: 77.33	Twitter sentiment dataset in SemEval 2013
Agarwal *et al.* [60]	Used mRMR approach for feature selection and then given input to SVM technique for classification	Support Vector Machine (SVM)	SVM: 90.1	Cornell movie review dataset and Amazon product review dataset
Foroozan *et al.* [61]	Considered n-gram approach with linear kernel based SVM and RBF kernel based SVM for classification	SVM with linear and RBF kernel	SVM: RBF: 92.1, Linear: 94	2308 news articles collected from Google finance.
Aisopos *et al.* [62]	Used supervised machine learning approach to test experimental results with multilingual manually annotated tweets from twitter	Multinomial NB (MNB), SVM, Logistic Regression (LR), C4.5 Tree, Multilayer Perceptron (MLP) and KNN	MNB:69.24, SVM: 70.4, LR: 70.36, C4.5: 70.33, MLP:70.11, KNN:70.2	95608 tweets combining corpora available to research community.

population of GA and also to different operations of GA, i.e., crossover and mutation. After completion of operations of GA, SVM with ten-fold cross-validation technique is used to classify the reviews. They use IMDb movie review dataset and data from two web forums, i.e., US supremacist for English language and a Middle Eastern extremist group for the Arabic language to perform the task of classification. They have obtained an accuracy of 91.7% for classification of the movie reviews, 90.6% for the data collected from US supremacist web forum in English, and 90.52% for the data collected from the Middle Eastern extremist group in Arabic.

- Zhao *et al.* have proposed topic modeling method which can automatically separate aspect and opinion words [64]. They have integrated a discriminant maximum entropy (MaxEnt) component with standard generative component. The MaxEnt component

allows to leverage features like POS tag to help separate aspect and opinion words. In their hybrid approach, they have run 500 iteration of Gibbs sampling and then used MaxEnt for classification. They have used three labeled dataset: one from restaurant dataset used by Ganu et al. [65] and other two hotel dataset used by Wu et al. [66] for finding aspect and opinion words and finally for classification.

- Feldman *et al.* have proposed a novel hybrid approach i.e., the stock sonar (TSS) to analyze the sentiment of the stocks [67]. The TSS integrates sentiment dictionaries, phrase level composition pattern and predicate level semantic events. It generates, precise in text sentiment tagging along with sentiment oriented event summarization for a given stock. TSS provides sentiment extraction that highlights positive and negative expressions within the article text. The extracted sentiments provide the analyst, an explanation for article score as well as summary for multiple news articles. They have developed a hybrid sentiment analysis approach, which combines three linguistics components for analysis, i.e., firstly a wide coverage sentiment lexicon, secondly patterns for modeling phrase-level compositional expressions and finally, semantic event extractor for business events. The TSS system produced a impact graph for Clinical Data Inc. from January 15th to January 27th, 2011 and based on the information obtained from this graph, 10,000 articles related to stock are collected for document level sentiment analysis.

- Govindarajan has proposed a sentiment classification technique using hybridization of NB and GA [68]. He has used the ensemble technique for sentiment classification. The combination of NB and GA is considered by the author as both techniques are highly heterogeneous on their approaches. He has used dimension reduction technique to reduce the size of the dataset. Best First Search (BFS) approach is used by him to select the best feature from the set of reduced feature. The text reviews are converted to vector form to give input to the hybrid of NB and GA, which use the voting mechanism to classify the reviews. The author has downloaded dataset from Bo Pang's web page which contains 2000 labeled movie reviews. The hybrid approach proposed by him shows an accuracy value to the extent of 93.8% on the labeled movie review.

- Basari *et al.* have proposed sentiment classification technique using the hybrid of SVM and Particle Swarm Optimization (PSO) technique [69]. The SVM-PSO approach is used to improve the accuracy of SVM by finding out the best features and regularization of the kernel of SVM. The PSO starts by choosing n-random particles and query for the optimal particle iteratively. PSO controls the possible subset selection mechanism for best prediction of accuracy. SVM uses the selected subsets and ten-fold cross validation technique to assess the performance. This task is carried out iteratively until the best possible accuracy value is obtained. They have used data collected from Twitter that can be found on the website with the URL address

as: ‘‘http://www.stanford.edu/ alecmgo/cs224n/trainingandtestdata.zip’’. The hybrid approach of SVM-PSO has shown an accuracy of 77% on the Twitter dataset.

- Agarwal and Mittal have proposed a sentiment classification technique using hybrid rough set based feature selection technique[70]. They have used rough set along with information gain (IG) to select the features. The IG based feature selection determines the important features from the documents. IG method does not consider the redundancy along attributes. Hence, it returns a large number of features for massive size dataset. They have used rough set attribute reduction technique to select the important features from the redundant set of features. They have considered polarity dataset [33] and reviews related to books, DVD, and electronics for classification. They have used SVM and NB technique for classification.

- Filho *et al.* have proposed a hybrid classification process with three different classification approaches, i.e., rule-based, lexicon based, and machine learning [71]. They have used a pipe-line approach for classification which works by back -off model. In their approach, each classifier classifies a review until certain level of confidence in accuracy is obtained. If the level is achieved, the final sentiment class is assigned to the review; else the review is provided to the next classifier. If still the level of accuracy is not obtained; then the voting mechanism is used for classification. The authors have used linear kernel based SVM for classification along with rule-based and lexicon-based approach. They have used five different datasets to test their proposed approach, i.e., Twitter2013, SMS2013, Twitter2014, LiveJournal2014, Twitter2014Sarcasm. Their proposed approach has shown an accuracy value to the extent of 53.31% and 65.39% on Twitter2013 and Twitter2014 dataset respectively.

- Khan *et al.* have proposed a hybrid approach for finding out the polarity of Twitter tweets [72]. They have suggested some pre-processing steps that include removal of URLs, hash-tags, special characters, substitution of abbreviations, and stop words removal. The classification algorithms used are the hybridization of enhanced emotional analysis, improved polarity classifier, and sentiwordnet analysis. Their proposed approach consists of three steps. First: data collection, i.e., the collection of twitter tweets based on some query; second: pre-processing of tweets and then transforming them into real value feature; and finally, use of the different classifier to classify the tweets. They have collected tweets based on six different searching criteria, viz., ‘‘Imran Khan’’, ‘‘Nawaz Sharif’’, ‘‘Dhoni’’, ‘‘Tom Cruise’’, ‘‘Pakistan’’, and ‘‘America’’ and then classified the tweets into different polarity groups. The accuracy is found to be of value to the extent of 88.89%, 82.86%, 86%, 85.55%, 85%, and 85.9% for dataset one to six respectively.

- Jagtap and Dhotre have proposed the sentiment classification of students' reviews regarding teachers feedback using hybridization of SVM and Hidden Markov Model (HMM) [73]. They have collected the data from the students regarding teachers' feed back. These review datasets are then processed to remove unwanted information and transformed them in the form of vector of real number for further machine learning use. These vectors are given input to both SVM and HMM separately for classification. After the classification, major voting rules are implemented to check as to which review to get how many votes to be a part of any polarity. In this voting mechanism both methods are used together. A max-min rule is applied later on to count the number of votes a review gets to be a part of positive polarity group and similar for negative reviews. All these counts are checked and at last, the reviews are classified into different polarity groups.

- Zhao and Jin have proposed a hybrid approach based on semantic labels which combines semantic based method with SVM technique [74]. They have considered each review text as a semantic phrase sequence and obtained two potential sentiment label for each review. They have assigned hybrid label as new feature to improve the performance of the approach. They have collected the reviews from Chinese movie review sites such as Mtive and DouBan movie. Their training set contains 2000 Chinese movie reviews while the testing set contains 1000 Chinese movie reviews.

- Nandi and Agrawal have proposed hybrid sentiment classification combining the lexicon dictionary based approach with SVM classifier result [75]. The lexical based approach depends on the dictionary of word i.e., bag-of-words for analysis and works on the principle that the polarity of the document is the sum of polarity of individual words or phrases. They have considered the twitter tweets for classification. They have collected the tweets related to Indian politics for classification.

- Desai and Mehta have proposed hybrid classification algorithms for analysis of students problems and perks [76]. They have combined both knowledge based and machine learning approach to process the tweet. They have collected the tweets with #engineringProblem as well as #engineeringPerks hashtags and considered them as the dataset for their analysis. In order to perform the knowledge-based analysis, a corpus is created with the collection of all the collected tweets. Then, the lexicons are found out with higher opinion values both for positive and negative polarity. These words are known as the seed words, and all possible synonyms and antonyms for these seed words are collected to form a lexicon dictionary. This lexicon dictionary is given input to the machine learning approach which considers it as an input and based on these input, the tweets are classified.

Table 2.3 provides a comparative study of different approaches adopted by various

authors in the area of sentiment classification using hybrid MLTS.

Table 2.3: Comparison of Sentiment Classification using hybrid MLTS

Author	**Approach Considered**	**Algorithm Used**	**Obtained result (Accuracy %)**	**Dataset used**
Abbasi *et al.* [63]	Used entropy weighted genetic algorithm along with SVM to classify the reviews	Hybrid of EWGA and SVM	Movie review: 91.7, US forum: 90.6, and Middle Eastern forum: 90.52	IMDb movie review dataset, and US supermacist and Middle Eastern extremist group web forums
Zhao *et al.* [64]	Integrated MaxEnt component with standard generative component and run the Gibbs sampling for 500 iteration to classify the reviews	MaxEnt with Gibbs Sampling	Restaurant dataset: 89.7, Hotel dataset: 82.0	Restaurant dataset used by [65] and two annotated hotel dataset used by [66]
Feldman *et al.* [67]	Proposed TSS system that uses three linguistic components, i.e., wide coverage sentiment lexicon, patterns for modeling phrase level and semantic event extractor for business events to classify the stocks	TSS system with linguistic components	TSS : 62.41	10000 articles collected from the graph generated for Clinical Data inc. from January 15 to January 27, 2011
Govindarajan [68]	Used hybrid of NB and GA to perform classification	Hybrid of NB and GA	Hybrid NB-GA: 93.8	Movie reviews downloaded from Bo Pang's web pages
Basari *et al.* [69]	Used hybrid of SVM and PSO technique to classify the reviews.	Hybrid of SVM and PSO	SVM-PSO: 77	The data collected from twitter
Agarwal and Mittal [70]	Used rough set based hybrid features selection technique to classify the reviews	SVM, NB	Movie: SVM:87.7, NB:80.9, Books: SVM:80.2, NB:79.1, DVD: SVM:83.2, NB:78.1,and Electronics: SVM:83.5, NB:78.1	Polarity dataset [33] and reviews collected for books, DVD and electronics
Filho *et al.* [71]	Used three different classification approaches, i.e., rule based, lexicon based, and machine learning based to classify the reviews.	Linear kernel SVM with rule based and lexicon based approach	Twitter2013 Dataset: 56.31, Twitter2014 dataset : 65.39	Twitter2013, SMS2013, Twitter2014, LiveJournal2014, Twitter2014Sarcasm
Khan *et al.* [72]	Used hybridization of enhanced emotional analysis, improved polarity classifier and Sentiwordnet analysis methods are used to classify the twitter tweets.	Enhanced Emotional Analysis, Polarity Classifier, and SentiWordnet analysis	Dataset1: 88.89, Dataset2: 82.86, Dataset3: 86, Dataset4: 85.55, Dataset5: 85, and Dataset6: 85.9	Tweets based on search conditions ''Imran Khan'', ''Nawaz Sharif'', ''Dhoni'', ''Tom Cruise'', ''Pakistan'', and ''America''
Jagtap and Dhotre [73]	Used SVM and HMM to classify the reviews independently and then using voting mechanism and max-min rule specify polarity to the reviews.	SVM and Hidden Markov Model (HMM)	SVM + HMM: 82.5	Reviews collected from student reading teachers performance
Zhao and Jin [74]	Have proposed semantic based method along with SVM for classification of reviews	Semantic approach with SVM	Hybrid approach: 80.1	Chinese movie reviews site, Mtime and DouBan Movie
Nandi and Agrawal [75]	Used combining lexical dictionary based approach along with SVM classifier for classification	Lexical dictionary based approach with SVM	Hybrid approach: 91	Twitter tweets based on Indian political issues
Desai and Mehta [76]	Used hybrid approach of knowledge based and machine learning based approach to classifying the twitter tweets.	knowledge based approach, machine learning approach (SVM)	Knowledge-based + SVM: 77.75	Twitter data of engineering students using #engineeringProblems and #engineeringPerks hash tags

2.5 Sentiment Classification using Feature Selection Mechanism

Sentiments are often classified based on text reviews, where each word of the text is considered as a feature. So, for a large size of the dataset, the number of features often tend to be very large. But among these features, it is observed that few are helpful for the sentiment classification. Thus, the task of selecting these subset of features from the set of all features is important for sentiment analysis. This Section discusses few kinds of literature that uses feature selection techniques for classification of the reviews.

- Neumann *et al.* have proposed four different feature selection techniques in order to improve the performance of the classifier [77]. They have used linear and non-linear SVM classifier for sentiment classification. The steps carried out by authors to perform feature selection are: firstly, improve the data collection process, reduce store space and classification time; secondly, they perform analysis of the reviews semantically to understand the problem and finally, improve the prediction accuracy of the classifier. The different feature selection techniques adapted by authors are filters, wrappers, embedded, and direct objective minimizing. Filter technique is most commonly used where the selection is carried out at pre-processing step only, irrespective of the classifier used. Wrappers mechanism takes help of the classifier as a black box component to perform the selection. Embedded method determines the features at the training time of classification. Finally, in direct objective minimization technique, the objective function is updated in a manner to select the best features. They have used several dataset from UCI repository [78] and Colon Cancer dataset from Weston *et al.* [79].

- O'keefe and Koprinska have evaluated a range of feature selectors and feature weights using both NB and SVM for classification of movie reviews [80]. They have used two new feature selection methods and three new feature weighting methods for analysis. The feature selection techniques used by them are: categorical proportional difference and two other methods based pn sentiment values, obtained from SentiWordNet (SWN) [81] are: SWN subjective score (SWNSS) and SWN proportional difference (SWNPD). Along with this, they have used three different feature weighting method such as feature frequency (FF), Feature presence (FP) and TFIDF for classification. They have used Polarity dataset [33] for classification which contains 1000 positively label and 1000 negatively label movie reviews for classification.

- Nicholls and Song have proposed a novel feature selection technique called document frequency difference (DFD) to identify words which are more useful in sentiment classification [82]. They have compared their proposed approach with three existing feature selection techniques i.e., χ^2, optimal orthogonal centroid and count difference.

They have considered the polarity dataset proposed by Pang and Lee [33] for classification and used Maximum entropy modeling technique for classification of these reviews.

- Wang *et al.* have proposed an effective feature selection technique based on Fishers' discriminant ratio [83] for text sentiment classification. They have used two kind of probability estimators, such as Boolean value and word frequency along with four other kinds of feature selection techniques , i.e., Information Gain (IG), Mutual Information (MI), Chi and Document Frequency (DF) for classification. They have designed two kinds of word sets as candidate feature set, i.e., one denoted by 'U' consisting of all words in the text and another denoted by 'I' consisting of words that appear in both positive and negative texts. They have used SVM technique for classification on two corpus, i.e., 2739 subjective documents of Chinese opinion analysis evaluation corpus belonging to different domains such as movie, education, finance, economics etc. and 1006 Chinese text reviews about eleven kinds of cars trademarked between January 2006 to March 2007.

- Maldonado *et al.* have proposed an embedded method of feature selection by penalizing each feature along with SVM classifier [84]. Their proposed approach is known as kernel-penalized SVM (KP-SVM) method. The KP-SVM method imposes stopping criteria to select best features from the set of features. The authors optimize the kernel function of SVM to eliminate the features that have negative effect on the performance of the classifier. This task is performed by gradient descent approximation of kernel function and feature selection. Their approach is best suited for SVM with RBF kernel. Four different datasets are used by the authors for the classification, i.e., two real words dataset from UCI repository and two DNA microarray dataset. To perform the classification, they have divided the dataset into training and testing groups. 60% of total observe data is considered for training and rest 40% are considered for training purpose.

- Sharma and Dey have proposed feature selection approach along with machine learning techniques to classify the movie reviews [85]. They have used five commonly used feature selection techniques, viz., Information Gain (IG), Document Frequency (DF), Gain Ratio (GR), CHI statistics, and Relief-F algorithm. Along with the feature selection techniques, they have also used five machine learning techniques, i.e., NB, SVM, ME, Decision Tree, KNN, Winnow, and AdaBoost for classification of the reviews. All the feature selection techniques compute a score for individual features and then a predefined number of features are selected as per ranking obtained from the score. They have used the polarity dataset for classification proposed by Pang and Lee [33]. The GR feature selection technique has shown the best result among all

other feature selection techniques and SVM shows the best result among the proposed machine learning techniques.

- Duric and Song have proposed a context and syntax model for feature selection [86]. They have considered only the subjective expression for analysis and ignored the expression that is less effective regarding polarity. Their feature selection mechanism is based on following concepts:

  i. *Features should be expressive enough to add useful information:* Those features are selected for analysis, which play an important role in providing information for sentiment classification.

  ii. *All features together should come from a broad and comprehensive viewpoint of entire corpus:* This signifies that from the set of selected features, it should be less difficult to find out the ideas that the document wants to express.

  iii. *Feature should be domain dependent:* In sentiment analysis each word is considered as a feature and a particular word may have different meaning in different domains thus the features need to be domain dependent.

  iv. *Feature should be frequent and discriminative enough:* The frequency of the words need to be higher, which have impact on the sentiment of the text and also are discriminative in nature, i.e., if the word is removed from the text, it affects the sentiment score of the text.

- Xia *et al.* have proposed a feature ensemble plus sample selection approach for domain adaption purpose [87]. The proposed approach has shown a significant improvement over feature ensemble (FE) and sample selection (SE) methods. They have collected features based on different POS tags. They have termed the feature with heavily changing POS tag as domain specific and the features with slight change in POS tag as domain independent. They have developed a feature selection method based on principal component analysis (PCA) as an aid to FE for feature selection. They have then used LDA technique to classify the reviews. They have collected reviews form Amazon.com based on four domains i.e., books, DVD, electronics, and kitchen for classification.

- Babatunde *et al.* have proposed the application of GA for feature selection purpose [88]. They have used binary-GA in order to reduce the dimensions and enhance the performance of the classifier. They have used k- nearest neighbor (KNN) based classifier error as the fitness function to find out fit chromosomes. In order to perform the GA analysis, the authors have transformed the text reviews into chromosomes. Then, using KNN, they have found the fitness value of each review. A selection criteria is adopted to select the best review. The reviews which are not fit, again go

though a process of GA operations. This process is repeated until required number of reviews are obtained. The authors have used five different datasets from Flavia dataset, i.e., Zernike moments, Lengendre moments, Hu 7 moment, Texture features, and Geometric feature. They use five different MLTs like multilayer perceptron (MLP), RF, j48, NB, and classification with regression module using Weka tool to perform the classification.

- Zheng *et al.* have proposed the feature selection for sentiment classification on Chinese review dataset [89]. They select ''N-char-gram'' and '' N-POS-gram'' approach to select the eligible sentiment features. Then they have improved document frequency method to select the feature subset and boolean weighting method is used to find the feature weight. They have used chi-square test to significantly improve the classification result. N-char-gram approach is adopted by the author as firstly, it does no require any POS tagging; secondly, it does not have any spelling error; and finally, no prior information is required for analysis. In N-POS-gram approach, mainly four POS tag words are considered, i.e., adjective, adverb, verb and noun. According to author, 4-POS-gram approach shows an better accuracy and the lower the value of 'N', i.e., number of consecutive characters in N-char-gram the result is better. They have collected mobile phone reviews from Jingdong, a famous electronics website which contains 1500 positive and negative reviews and use SVM technique to classify the reviews.

- Agarwal and Mittal have initially extracted features like unigram, bigram, and dependency features from text [90]. They have then extracted bi-tagged features to confirm as per POS pattern. Then information gain (IG) and maximum redundancy maximum relevancy (mRMR) feature selection mechanism is used to eliminate irrelevant features from the feature vector. Redundancy among the features is reduced by mRMR feature selection mechanism. IG is measured by reducing the uncertainty in identification of the class attributes when the value of the features is known. Boolean Multinomial NB and SVM machine learning techniques are used to classify the reviews. They have used Cornell movie review dataset and Amazon product review dataset to perform the classification process.

- Uysal has proposed an improved global feature selection scheme (IGFSS) at the end of the feature selection mechanism to obtain more representative feature set [91]. The effectiveness of IGFSS is accessed using local feature selection mechanism like Information Gain (IG), Gini Index (GI), Distinguishing Feature Selector (DFS) and odds ratio (OR). IG score indicates the ratio of presence or absence of any term to correctly classify a text document. GI defines an improvement over the attribute selection algorithm used in decision tree approach for feature selection. DFS selects the discriminative features that removes the irrelevant ones using some predefined

condition. OR matrix measures the belongingness or not belongingness to a class with nominator and denominator respectively. He has used both SVM and NB technique for classification. For the classification purpose, he has used top 10 classes of the celebrated Reuters-21578 ModApte split, WebKB dataset consist of four classes, and Classic3, whose class distribution is nearly homogeneous among three classes.

Table 2.4 provides a comparative study of different approaches adopted by authors, contributed to sentiment classification using feature selection technique.

Table 2.4: Comparison of Sentiment Classification using feature selection techniques

Author	Approach Considered	Algorithm Used	Obtained result (Accuracy %)	Dataset used
Neumann *et al.* [77]	Used four different feature selection techniques continuously with linear and non-linear SVM to perform classification	Liner and non-linear SVM	SVM: 90	several dataset from UCI repository [78] and Colon Cancer dataset from Weston et al. [79].
O'keefe and Koprinska [80]	Used feature weights to select features and use both NB and SVM for classification	NB and SVM	SVM: FF:85.5, FP:87.15, TF-IDF:86.55, NB: FF: 77.2, FP: 81.5, TF-DF: 77.6	Polarity dataset proposed by Pang and Lee [33]
Nocholls and Song [82]	Used feature selection technique document frequency difference (DFD) to select features and then used ME modeling technique to classify the review	Maximum Entropy Modeling	ME: f-value :79.9	Polarity dataset proposed by Pang and Lee [33]
Wang *et al.* [92]	Used feature selection technique based on Fishers' discriminant ratio [83] and then used SVM for classification	SVM	SVM: 86.61	2739 reviews collected from Chinese Opinion Analysis Evaluation corpus and 1006 Chinese reviews about eleven kinds of car
Maldonado *et al.* [84]	Used embedded method of feature selection along with SVM RBF kernel to perform classification.	SVM	DIA: 76.54, WBC: 97.5, CMA: 96.5, and LMA: 99.7	Diabetes data set (DIA), Wisconsin Breast Cancer (WBC), Colorectal Microarray data set (CMA), and Lymphoma Microarray data set (LMA).
Sharma and Dey [85]	Used five feature selection techniques along with seven machine learning algorithm to classify the reviews.	NB, SVM, ME, DT, KNN, Winnow, and AdaBoost	NB: 90.90, SVM: 90.15, ME:88.85, DT:75.35, Winnow: 73.3, AdaBoost: 66.90	Polarity dataset proposed by Pang and Lee [33]
Duric and Song [86]	Used a context and semantic model for feature selection along with HMM-LDA method for classification of reviews.	HMM and LDA	10 iteration: 82.3, 25 iteration: 83.9, and eval: 86.3	2000 movie reviews dataset used by Pang and Lee [33]
Xia *et al.* [87]	Proposed a feature ensemble plus sample selection technique along with LDA to classify the reviews	LDA	LDA: 84.87	Reviews collected from Amzon.com based on four different domains books, DVD, electronics, and kitchen
Babatunde *et al.* [88]	Used binary GA and KNN based classification error as fitness function to select the best feature. Then used different MLTs in Weka tool to perform the classification.	multilayer perceptron (MLP), RF, j48, NB, and classification with regression tool	Information gain + MLP: 93.73	five different datasets from Flavia dataset, i.e., Zernike moments, Lengendre moments, Hu 7 moment, Texture features, and Geometric feature
Zheng *et al.* [89]	Used n-char-gram and N-POS-gram technique for feature selection then use SVM to classify reviews.	SVM	4-POS-gram with 150 features: 96.52, 1-char-gram with 150 features: 95	Chinese mobile phone reviews from Jing Dong, a famous electronics e-commerce site.
Agarwal and Mittal [90]	Extracted features using unigram, bigram and dependency feature technique, them use NB and SVM to classify the reviews.	Boolean Multinomial NB, SVM	Boolean Multinomial NB: 91.8	Cornell Movie review Dataset, Amazon product review Dataset.
Uysal [91]	Used global feature section scheme along with local feature section techniques for selection of feature and then use MLTs to classify the review.	SVM and NB	Reuters dataset: SVM: 67.076, NB : 68.514, WebKB dataset: SVM:83.4, NB:84.78, and Classic3 dataset: SVM: 97.9, NB: 99.02	celebrated Reuters-21578 ModApte split dataset, WebKB dataset, and Classic3 dataset

2.6 Sentiment analysis using unsupervised machine learning approach

The Section 2.2, Section 2.3, Section 2.4, and Section 2.5 discuss about the sentiment analysis using supervised machine learning approach. In this approach, the dataset needs to be a labeled one. But, collecting labeled data from a reliable source is a difficult task. On the other hand, it is easier to obtain the unlabeled data. The analysis of these unlabeled

data is carried out using unsupervised machine learning approach. This section discusses on few of this category of articles, where unsupervised MLTs approach is used. They are highlighted as below:

- Kanayama and Nasukawa have proposed the Japanese version of domain oriented sentiment analysis [93]. The proposed approach selects the polarity clauses conveying the goodness and badness in specific domain. For lexicon based analysis, they have used unlabeled corpus and have assumed that polar clauses with same polarity appears successively unless the context changed with adversative expressions. Using the approach, they have collected candidate polar atoms and their probable polarities. They have also considered inter-sentential and intra-sentential context to obtain more polar atoms. They have also found out coherent precision and coherent dependency, which help to analyze document in new domain. They have collected Japanese corpora from discussion board based on four different domains, i.e., digital cameras, movies, mobile phones and cars for unsupervised sentiment analysis.

- Wan has considered the Chinese reviews for unsupervised sentiment analysis as it is difficult to obtain labeled reviews for analysis [94]. He has translated the Chinese reviews into English reviews using Google Translator, Yahoo based Fish and Baseline translator. He has then used ensemble methods to combine the individual analysis results of both the language to improve analysis result. He has used six different ensemble techniques, i.e., average, weighted average, max, min, average of max and min and majority voting for unsupervised sentiment analysis. He has collected 1000 product reviews from a popular Chinese IT product web site-IT168.

- Zagibalov and Carroll have proposed automatic seed work selection for the unsupervised sentiment classification of Chinese reviews [95]. They have initially considered a single human selected word ‘good’ for analysis and used a iterative method to extract a training sub-corpus. They have used the term ‘‘lexical item’’ to denote any sequence of Chinese characters and ‘‘zones’’ to represent the sequence of characters finished by punctuation marks. Each zone is then classified into different polarity groups based on predomination of polarity vocabulary items. In order to find out the polarity of a document, the difference between the positive and negative zones are carried out and if the difference is found out to be positive, then the document is classified as positive else negative. They have considered the product reviews obtained from IT168 website for classification which contains 29531 reviews after removing duplicate reviews.

- Rothfels and Tibshirami have examined the unsupervised system by iteratively extraction of positive and negative sentiment items from text, then classified the document based on these information [96]. They have worked on the principle of

positivity of language i.e., most of the words associated with sentiment extracted through unsupervised approach are found to be positive. In order to collect the positive words, they have first extracted all adverbial phrases of set lengths that follows negative words, and then pruned these words in to the list of existing corpus. They have again validated the sentiment of the text by capturing the small lexical units i.e., adjective and adverbial phrases. They have used Polarity dataset proposed by Pang and Lee [33] for classification of reviews.

- Lin *et al.* have compared three closely related Bayesian models, i.e., latent sentiment model (LSM), joint sentiment topic model (JST), and Reverse JST model for unsupervised sentiment classification [97]. The LSM model is a combination of three different labels i.e., positive, negative, and neutral. The JST model can detect both sentiment and topic simultaneously by modeling each document with topic document distribution. Reverse JST is a four layered hierarchical Bayesian model where topics are associated with document under which words are associated with topics and sentiment labels. They have used MR dataset proposed by Pang *et al.* [8], subjective MR dataset proposed by Pang and Lee [33] and dataset containing four different types of product reviews collected from Amazon.com [53] for classification.

- Paltoglou and Thelwall have focused on textual communication available on web, i.e., through Twitter, MySpace and Digg as they are less domain specific and unsupervised in order to make polarity prediction [98]. Their approach consists of two different contexts, i.e., subjectivity detection and polarity classification. They have added a list of linguistically driven functionalities to the classifier such as negation / capitalization detection, intensifier detection, and emotional / exclamation detection to help in final prediction. They have used three different datasets from real world for analysis. They have collected dataset form Twitter i.e., divided into two subsets, i.e., tweets collected from Twitter API and humanly annotated tweets. Second dataset is collected from Digg, from February to April 2009 and contains 1.6 million tweets. Finally, third dataset is collected from MySpace for analysis.

- Ghosh and Kar have proposed a pattern based method by applying classification rule using unsupervised machine learning approach [99]. They have used SentiWordNet to calculate sentiment score of each document. They have finally combined the sentiment score of each sentence to predict the sentiment of the document. They have used SentiWordNet to assign each synset of the WordNet with three sentiment scores, i.e., objective, positive and negative. They have ignored the objective sentences during the classification process. They have considered reviews related to three kinds of digital cameras i.e., Canon EOS40D, Nikon Coolpix, and Nikon D3SLR which are collected from Amazon.com and ebay.com.

- Hu *et al.* have proposed a study of unsupervised sentiment analysis with emotional signals [100]. They have modeled two main categories of emotional signals i.e., emotion indication and emotion correlation. They have represented emotional signals with statistical hypothesis testing and proposed a unified way to model the emotional signals. They have also used emotional indication which strongly reflect the sentiment polarity of a post or a review. They have proposed emotional correlation which reflect correction between two words as per emotional consistency theory [101]. They have collected tweets from Standford Twitter sentiment form April 6, 2009 to June 25, 2009 using Twitter API and Obama-McCain debate dataset [102] which contains tweets during presidential debate on September 26, 2008 for classification.

- Milagros *et al.* have proposed an approach to predict sentiment of tweets and reviews based on unsupervised dependency parsing based text classification method [103]. Their approach leverages a variety of NLP and sentimental features primarily derived from sentiment lexicon. They have used two different variants of their approach: the sentiment lexicon with polarity rank 40 (PR 40) created with a total number of 40 positive and negative seeds and sentiment lexicon SO-CAL with PR40. They have used three different dataset to test their approach. They have used Cornell movie review dataset as proposed by Pang and Lee [33], Obama-McCain debate dataset [102] and SemEval -2015 task 10 dataset for classification.

- Biagioni has proposed unsupervised method of sentiment analysis [104]. His proposed approach consists of two components such as bespoke sentiment analysis system developed by author and SenticNet sentiment lexicon. The sentiment lexicon acts as the source of sentiment information and performed the sentiment classification task. SenticNet defines the technique for identifying concepts and retrieving their polarity values as well as threshold value, used to submit two single word and multi-word classification strategy. He has considered two datasets from two different domains for analysis. The first dataset is emotion related dataset obtained from International Survey on Emotion Antecedents and Reaction (ISEAR) project and polarity dataset as proposed by Pang and Lee [33].

Table 2.5 provides a comparative study of different approaches adopted by authors, contributed to sentiment classification using unsupervised approach.

Table 2.5: Comparison of Sentiment Classification using unsupervised approach

Author	**Approach Considered**	**Algorithm Used**	**Obtained result (Accuracy %)**	**Dataset used**
Kanayama and Nasukawa [93]	Used Japanese version of domain oriented sentiment analysis and select polarity clause and based on that perform the classification	classify reviews based on polarity atoms and lexical rule	Precision: Digital Cameras: 96.5, Movies: 94.4, Mobile phones: 92.1, Cars: 91.5	Japanese corpora from discussion board based on four different domains, i.e., Digital Camera, Movies, Mobile Phones, and Cars.
Wan [94]	Translate the Chinese reviews to English using machine learning translators and then use ensemble methods to classify the reviews	Ensemble Techniques	Average: 85.4, Weighted Average: 86.1, Max: 82.3, Min: 84.8, Average of max and min: 84.3, Majority voting: 82.3	1000 product reviews from a popular Chinese IT product web site-IT168
Zagibalov and Carroll [95]	Identify lexical item and Zone, then sentiment value of the zones are found out and based on that classification carried out.	Lexicon based analysis	Highest F1 value : 89.91	Chinese product review dataset obtained from IT168, consist of 29531 reviews
Rothfels and Tibshirami [96]	Iteratively extract positive and negative sentiment words from text and based on that classify the reviews.	classification based on adjective and adverb selection.	Accuracy: 65.5	Polarity dataset proposed by Pang and Lee [33]
Lin *et al.* [97]	Compared three Bayesian model for unsupervised document level classification.	latent sentiment model (LSM), joint sentiment topic (JST) model, reverse JST model.	MR: LSM: 74.1, JST: 70.2, RJST: 68.3, Subjective MR: LSM: 57.9, JST : 73.4, RJST : 71.2, MDR: LSM: 73, JSR : 66.1, RJST: 65.3	MR dataset proposed by Pang *et al.* [8], subjective MR dataset proposed by Pang and Lee [33] and dataset containing four different types of product reviews collected from Amazon.com [53]
Paltoglou and Thelwall [98]	Used Subjective detection and polarity classification technique to classify the reviews.	SVM, NB, ME	Twitter: SVM: 73.2, NB:75.9, ME:73.3, MySpace: SVM : 73.2, NB: 72.6, ME: 63.6, Digg: SVM: 72.7, NB: 69.2, ME:70.1	Textual reviews collected from web through Twitter, MySpace and Digg.
Ghosh and Kar [99]	Use SentiWordNet to calculate sentiment score of each word then combine them to classify the reviews.	Sentiment score obtained from SentiWordNet	F1 measure: Canon: 86.7, Nikon Coolpix: 79.79, Nikon D3SLR: 84.47	Consider reviews related to three kinds of digital camera, Canon EOS40D, Nikon Coolpix, NIKON D3SLR collected form Amazon.com and ebay.com.
Hu *et al.* [100]	performed unsupervised sentiment classification using emotional signals.	Emotional Signals for unsupervised Sentiment Analysis (ESSA).	ESSA: STS: 73.5, OMD: 68.6	Tweets from Standford Twitter sentiment form April 6,2009 to June 25,2009 using Twitter API and Obama-McCain debate dataset [102] which contains tweets during presidential debate on September 26, 2008.
Milagros *et al.* [103]	Extract sentiment lexicons from the tweets and reviews and based on that perform the classification.	PolarityRank 40 (PR40), Sentiment lexicon with PolarityRank 40 (SO-CAL + PR40)	Cornell : PR40: 68.6, SO-CAL + PR40: 69.95, OMD: PR40: 70.75, OMD: 71.3	Cornell Movie review Dataset [33], Obama-McCain debate dataset [102]
Biagioni [104]	Used bespoken sentiment analysis approach and SenticNet sentiment lexicon approach for unsupervised sentiment classification.	SenticNet sentiment lexicon	Accuracy: 73.4	emotion related dataset obtained from International Survey on emotion antecedents and reaction (ISEAR) project

2.7 Sentiment analysis using semi-supervised machine learning approach

The Section 2.2, Section 2.3, Section 2.4, and Section 2.5 discuss about the sentiment analysis using supervised machine learning approach. Section 2.6 is concerned about the sentiment classification using unsupervised approach. This section discusses about sentiment analysis using semi-supervised approach. In the present day scenario, a small size of labeled dataset is present where the size of unlabeled dataset is large. Based on

the information obtained from the labeled dataset, the unlabeled dataset is transformed in to labeled one [31]. This section discusses on few of these categories of articles where semi-supervised MLTs approach is used. They are highlighted as below:

- Goldberg and Zhu have proposed a graph based semi-supervised learning approach to address the sentiment analysis task of rating interference [105]. They have created graph on both labeled and unlabeled data to encode certain assumptions for the task. They have then solved an optimization problem to obtain a smooth rating function of overall graph. They have assumed that the similarity measure between two documents should be greater than equal to zero. They have performed the experiment with positive sentence percentage and mutual information modulated word vector cosine similarities. They have considered the movie reviews documents accompanying four different class labels found in ‘‘Scale dataset v1.0’’ available at Cornell digital library [106].

- Sindhwani and Melville have proposed a semi supervised sentiment prediction algorithm which utilizes lexical prior information with unlabeled examples [107]. Their method is based on joint sentiment analysis of document and words based on a bipartite graph representation of the data. They have incorporated sentiment laden terms to their model for analysis. In order to adopt to a new domain with minimal supervision, they have exploited large amount of unsupervised data. They have used movie reviews dataset proposed by Pang and Lee [33], along with two other blog dataset for analysis. They have created a dataset targeting the detection of sentiment, which contains information about IBM Lotus brand. The second blog dataset consist of 16742 political blogs.

- Melville *et al.* have presented a unified framework that uses background lexicon information in terms of word-class association and refines the information for specific domain [108]. They have constructed a generative model based on lexicon of sentiment laden words and another model to train the label dataset. These two models are adaptively pooled to create a composite multinomial NB classifier to capture information. They have used the labeled document to refine the information collection which is based on generic lexicon effective for all domains. They have considered 20488 technology blogs which contains 1.7 million post from IBM Lotus collaborative software, 16,741 political blog containing two million posts and the movie review dataset proposed by Pang and Lee for classification.

- Lazarova and Koychev have proposed a semi-supervised multi-view learning approach for sentiment analysis in Bulgarian language [109]. They have considered 992 English labeled examples from Amazon.com and TripAdvisor.com, and then translated them into Bulgarian for analysis. They have also extracted a set of 100

movie reviews in Bulgarian from *www.cinexio.com* and translated it to English using on-line translation software to check the accuracy while translating from Bulgarian to English and vice versa. They have used semi supervised multi-view Genetic Algorithm (SSMVGA) for analysis of reviews both in English and Bulgarian. They have compared their result with supervised approach on both English and Bulgarian dataset using Root Mean Square Error (RMSE) approach and found that their proposed approach has shown the better result.

- Anand and Naoream have proposed semi supervised aspect based sentiment analysis (ABSA) approach to classify movie reviews [110]. Their approach consists of two parts. They have considered two class classification scheme for plots and review without considering the labeled data. Secondly, they have considered a scheme to detect aspects and corresponding opinion using manually crafted rules and aspect clue words. They have considered three different schemes for selection of aspect clue words, i.e., manual labeling, clustering and review guided clustering for selection of aspect words from the reviews. They have filtered the sentimets from reviews, then extracted sentiments from these reviews and then associate it with corresponding aspect category. They have considered Amazon movie review dataset for analysis. The dataset contain 7911684 reviews provided by 889176 users on 259059 movies / TV shows.

- Miyato *et al.* have proposed a virtual adversarial training for semi-supervised text classification. They have used extended adversarial and virtual adversarial training to the text domain by applying perturbations to the word embedding in a recurrent neural network. They have used neural language model proposed by Dai and Le [111] to achieve multiple semi supervised text classification tasks including sentiment classification and topic classification. They have initialized word embedding matrix and LSTM weight with pre-trained recurrent language model which train both labeled and unlabeled reviews. Based on this training information, the testing of other reviews is carried out. They have used five different datasets i.e., IMDb dataset [32], Elec an Amazon electronics product review dataset [112], Rotten Tomatoes consist of small snippet of movie reviews [106], DBpedia a dataset of Wikipedia pages [113], and RCV1 dataset consist of news articles from Reuters corpus [114] for classification.

- Silva *et al.* have presented a survey of tweet sentiment analysis using semi supervised learning technique [115]. They have identified three categories of semi-supervised approach namely graph based, topic based method, and wrapper based (self training and co training), for tweet sentiment analysis. The graph-based methods propagate labels to unlabeled data. The label propagation process requires the computation of similarities among the data instances. The topic based approach captures local information from the data, i.e., unigram, bigram, POS tag, and lexicon information

from the dataset and based on that, it performs the classification. They have used widely known semi supervised learning (SSL) approach namely Self training [116] and co-training [117] for classification of tweets based on the F1 values.

- Khan *et al.* have incorporated machine learning approach with lexical based approach and introduced a new framework called Semi supervised feature weighting and intelligent model selection (SWIMS) to determine feature weight based on general purpose sentiment lexicon SentiWordNet [118]. They have used SVM to learn feature weights and applied an intelligent model selection approach to enhance the classification performance. For feature selection, they have used different POS, point wise mutual information and Chi-square test. They have also used SentiWordNet lexical resource that contains polarity value for feature selection. They have considered seven different datasets to test their approach. The dataset are: Large movie review dataset consist of 50000 movie reviews [32], Cornell movie review dataset proposed by Pang and Lee [33] and five dataset, i.e., Apparel, Books, DVDS, Health and Video considered from multi-domain sentiment dataset [53].

- Wang *et al.* have constructed a novel kernel eigenvector by injecting the class label information under the framework of eigenfunction extrapolation [119]. They have designed a base kernel used in semi supervised kernel learning. Besides using the eigenvector from kernel matrix, they have computed a new set of eigenvectors which are expected to be better aligned to the target. They have used class labels to improve the quality of base kernels using framework of eigenfunction extrapolation, links between class labels and ideal kernel eigenfunction. They have extended the approach to multiple kernel setting to improve the modeling power of proposed approach and finally compared the state-of-art semi supervised approach under single and multi-kernel setting.

- Da Silva *et al.* have proposed semi supervised based learning (SSL), which combines unsupervised information collected form similarity matrix constructed from unlabeled data [120]. They have integrated clustering ensemble (C^3E) [121] algorithm with SSL framework for classification of Twitter tweets. They have combined the SVM classification information collected from labeled data with information obtained from pair-wise similarity between unlabeled data points. Their proposed framework based on iterative self-training approach is guided by predictions. The C^3E algorithm combines classification and clustering algorithms to obtain a better classification result. They have considered six datasets for analysis. They have collected dataset SemEval 2013 from international workshop on semantic evaluation (SemEval) by combining SemEval 2013 (Task 2) and SemEval 2014 (Task 9). They have also considered five different datasets namely LiveJournal[122], SMS2013 [123], Twitter2013 [123], Twitter2014 [122], and Twitter Sarcasm 2014 [122] for analysis.

Table 2.6 provides a comparative study of different approaches adopted by authors, contributed to sentiment classification using semi-supervised approach.

Table 2.6: Comparison of Sentiment Classification using semi-supervised approach

Author	**Approach Considered**	**Algorithm Used**	**Obtained result (Accuracy %)**	**Dataset used**
Goldberg and Zhu [105]	Created graph of labeled and unlabeled dataset, then solve an optimization problem to obtain smooth rating of graph	Positive sentence percentage and mutual information modulated word vector cosine similarities	Semi-supervised learning + PSP : 68.9	Movie review document accompanying 4 class labeled available in Cornell digital library and first used by Pang and Lee [106]
Sindhwani and Melville [107]	Used lexicon prior knowledge and joint sentiment analysis of document along with word based bipartite graph to predict the sentiment of the text.	Lexical information (LEX), Regularized least square algorithm (RLS), and Semi-supervised lexicon with RLS (SSL+RLS)	Movie: LEX + RLS: 72, SSL+RLS: 75, Political: LEX + RLS: 61, SSL+RLS: 65, IBM Lotus: LEX + RLS: 88, SSL+RLS: 93	Movie review dataset [33], 14 individual blogs collected from IBM Lotus brand, 16742 Political blogs.
Melville *et al.* [108]	Used background lexical information in term of word-class association and refine the information for specific domain for analysis.	Lexical classifier (LC), Feature supervision (FS), NB, Linear pooling (LP), Log pooling (LOP)	Movie: LC:63.4, FS:57.59, NB: 80.81, LP: 81.42, and LOP: 80, Politics: LC: 55.2, FS: 46.19, NB: 59.24, LP: 63.61, and LOP: 60.04, IBM Lotus: LC:68.23, FS:57.93, NB: 88.40, LP: 91.21, and LOP: 88.42	20488 technology blogs which contain 1.7 million post from IBM Lotus collaborative software, 16741 political blog containing 2 million post and the movie review dataset proposed by Pang and Lee [33].
Lazorova and Koychev [109]	Used multi-view approach for sentiment analysis in Bulgaria language for sentiment analysis.	Semi-supervised multi-view Genetic Algorithm (SSMVGA) using Root Mean Square Error (RMSE)	SSMVGA : RMSE: 2.16	992 labeled English reviews and 100 movie reviews in Bulgarian from www. cinexiv.com
Anand and Naorem [110]	Used aspect based sentiment analysis approach to classify the movie reviews.	Manual labeling (M), Kerword Clustering (KC), and Keyword Review filtered clustering (KRC)	M: 72, KC: 68, KRC:70	Amazon movie review dataset consist of 7911684 reviews provided by 889176 users on 259059 movie or TV shows
Miyato *et al.* [124]	Used extended adversarial and virtual adversarial training on text domain with recurrent neural network for text classification.	Neural Network expressing test error rate	Test error rate: IMDB: 5.91, Elec: 6.24, Rotten Tomatoes: 19.1, DBpedia: 3.57 and RCV1: 8.52	IMDb dataset [32], Elec an Amazon electronics product review dataset [112], Rotten Tomatoes consist of small snippet of movie reviews [106], DBpedia a dataset of wikipedia pages [113], and RCV1 dataset consist of news articles from Reuters corpus [114]
Silva *et al.* [115]	Used widely known semi-supervised sentiment learning approach i.e., self training [116] and co-training [117] for classification	Self training and co-training approach	Self training: Livejournal: 70, SMS2013: 65, Twitter2013: 65, Twitter2014: 50, Twitter Sarcasm : 80, Co-training: Livejournal: 30, SMS2013: 55, Twitter2013: 35, Twitter2014: 50, Twitter Sarcasm : 20,	Five test datasets namely LiveJournal, SMS2013, Twitter2013, Twitter2014, and Twitter Sarcasm 2014
Khan *et al.* [118]	Incorporated machine learning approach with lexical based approach and introduced the frame work Semi supervised feature weighting and intelligent model selection (SWIMS) to select feature. Then use SVM and intelligent model selection approach to classify the reviews.	SWIMS with 10 fold (SWIMS10) and SWIMS with intelligent model selection (SWIMSIMS)	SWIMS10 : Cornell: 83.4, Large movie dataset: 85.96, Apparels: 81.05, Book: 77, DVD: 78.55, Health : 78.35, Video: 80.9, SWIMSIMS: Cornell: 85.5, Large movie dataset: 86.44, Apparels: 84.05, Book: 81.5, DVD: 81, Health : 81, Video: 82	Large movie review dataset consist of 50000 movie reviews [32], Cornell movie review dataset proposed by Pang and Lee [33] and five dataset, i.e., Apparel, Books, DVDS, Health and Video considered from multi-domain sentiment dataset [53].
Wang *et al.* [119]	Constructed a kernel eigenvector by injecting class label information, then extend it into multiple kernel system to obtain the classification result	Semi supervised learning with single kernel (SSLS) and Semi supervised learning with multiple kernel (SSLM)	SSLS: 86, SSLM : 91	Cornell movie review dataset [33]
Da Silva *et al.* [120]	Integrated clustering ensembles C^3E algorithm with semi supervised learning framework for classification of twitter tweets.	F1 measure with SVM classification	F-value: LiveJournal: 65.37, SMS2013: 54.9, Twitter2013: 57.13, Twitter2014: 56.68, and Twitter Sarcasm 2014: 45.54	SemEval2013 along with five different datasets namely LiveJournal[122], SMS2013 [123], Twitter2013 [123], Twitter2014 [122], and Twitter Sarcasm 2014 [122]

2.8 Summary

This chapter presents various sentiment classification techniques proposed by different authors for better classification result. However, there are few shortcomings present in these approaches. In next four chapters, the steps are taken to remove the shortcomings and to obtain better result after classification. It is observed from this chapter that finding a right kind of dataset is one of the big concerns and a good number of authors have preferred to consider the movie review datasets proposed by Pang and Lee [33]. Again it is observed that, MLTs are used by different authors for classification purposes which take the numerical values as input. The process of conversion of text reviews into numerical values is also a big concerned. Therefore, in subsequent chapters an attempt has been made to classify review dataset on movies, using different machine learning techniques with an aim to study about the improvement in the result after classification.

Chapter 3

Classification of Sentiment of Reviews using Supervised Machine Learning Techniques

This chapter presents a study on sentiment classification technique for review of a particular type of dataset i.e., on movie. First, a brief information about sentiment classification is provided. Then different methodologies used for classification are described. Afterwards, the proposed approach for classification is discussed and finally, different performance evaluation parameters are considered to assess the performance of various classifiers considered. Then, the obtained result is compared with the results obtained in existing literature in order to check the validity of proposed approach.
The rest of the chapter is organized as follows:
Section 3.1 provides a brief introduction to the proposed approach and the contribution of the chapter. Section 3.2 indicates the motivation for the proposed approach. Section 3.3 discusses about different techniques to transform text data into numerical vectors, classification techniques, and performance evaluation parameters. Section 3.4 highlights the proposed approach. Section 3.5 compares the performance of the proposed approach with present literatures. Finally, Section 3.6 summarizes the chapter.

3.1 Introduction

Sentiment analysis deals with study of people's opinion or review about any topic or product and provides a meaningful information on the topic. In order to analyze these reviews, different machine learning techniques are considered. In order to evaluate the performance of these techniques, different performance evaluation parameters are used. The contributions of this chapter can be explained as followed:

i. Two different movie review datasets i.e., IMDb [32] and Polarity [33] are considered for classification. The IMDb dataset contains separate dataset for both training and testing; whereas polarity dataset does not have separate dataset for training and testing. Thus, 10 fold cross validation technique is used for classification in polarity dataset.

ii. Four different machine learning techniques, viz., Naive Bayes (NB), Support Vector Machine (SVM), Random Forest, and Linear Discriminant Analysis (LDA) have been considered for classification on both datasets.

iii. Different performance evaluation parameters, i.e., precision, recall, f-measure and accuracy based on elements from confusion matrix are used to evaluate the performance of the MLTs.

3.2 Motivation for the proposed approach

The Section 2.2 discusses about document level sentiment analysis using different MLTs and the Table 2.1 provides a comparative analysis of those papers. These information help to identify some possible research opportunities which can be extended further. The following aspects have been considered for carrying out study on SA.

i. Most of the authors have preferred to validate their approach on a single dataset. In this chapter, more than one number of dataset are considered for classification. The two datasets i.e., IMDb and Polarity, are considered in such a manner that the approach for classification is different in both cases.

ii. A good number of authors have used NB and/or SVM techniques for classification. In case of NB, a good number of authors have considered one version, whereas there are three different versions of NB technique, i.e., Gaussian NB, Multinomial NB and Bernoulli NB. In this chapter, all three versions are implemented. SVM is a kernel based classification technique and most of the authors have used only liner kernel based SVM technique. In this chapter, different versions of kernels are taken into consideration, i.e., linear, polynomial, Gaussian radial basis function, and sigmoid for classification in order to identify to which one yields the best results.

iii. Most of the authors have used NB and / or SVM technique for classification of reviews. NB uses probabilistic Bayesian method for classification, whereas SVM uses kernel based system for classification. Thus, these two techniques are used for classification and along with that, in this chapter, two other MLTs are proposed, i.e., Random forest (RF) and Linear Discriminant Analysis (LDA). RF uses an ensemble method where weaker models work independently and their result is combined to obtain the final result. On the other hand, LDA uses a discriminant analysis method that creates linear combination of dependent variable based on independent variables and classifies the reviews.

3.3 Methodology Adopted

This section discusses about the different methodologies adopted for classification of sentiment reviews.

3.3.1 Types of sentiment classification

Sentiment classification process is mainly of two types, which are as follows:

(i) Binary sentiment classification: In binary classification each document d_i in D, where D = {d_1, d_2,….,d_n} is classified as a label C, where C is a predefined category set as C = {Positive and Negative}.

(ii) Multi class sentiment classification: In multi class sentiment analysis, each document d_i is classified as a level in C^*, where C^*= {strong positive, positive, neural, negative, and strong negative}.

Generally the binary classification is useful when the comparison between two products is done or when solving a two class problem. In this chapter, analysis based on binary sentiment classification has been carried out.

3.3.2 Transformation of Text Data into Numerical values

SA deals with reviews in the text format and MLTs are used to classify these reviews. But the MLTs do not process the text data. Thus, they need to be converted into numerical values or arranged in a form of matrix of numbers, which the MLTs take as input for both training and predicting the polarity of review. The different functions used to transform the text data into numerical values are explained as below:

- **CountVectorizer (CV):** It converts the text document collection into a matrix of token counts [125]. This function generates a sparse matrix of the counts. The following example shows, how the CV matrix is generated.

  Suppose; there exist a document containing following sentences.

  ‘‘ This car is speedy. ’’

  ‘‘ This car is beautiful. ’’

  ‘‘ This car is dirty. ’’

  A CV matrix of size 3*6 is generated using above sentences, because there exists 3 documents and 6 distinct features. Table 3.1 displays the matrix of numerical data for this case.

Table 3.1: Example of CV matrix

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6
Sentence 1	1	1	1	1	0	0
Sentence 2	1	1	1	0	1	0
Sentence 3	1	1	1	0	0	1

It can be observed from the Table 3.1 that the presence of the feature is marked by '1' while the absence of these are marked by '0'. The ''Sentence 1'' contains the first four words/ features thus these are marked as '1', while for ''Sentence 2'' and ''Sentence 3'' the ''Feature 4'' is marked as '0' while ''Feature 5'' and ''Feature 6'' are marked as '1' respectively.

- **Term Frequency Inverse Document Frequency (TF-IDF):** TD-IDF reflects the importance of a word in the corpus or the collection [125]. TF-IDF value increases with increase in frequency of a particular word that appears in any document. In order to control the generality of more common words, the term frequency is offset by the frequency of words in corpus. Term frequency is the no. of times a particular term appears in the text. Inverse document frequency measures the occurrence of the term in all documents. The TF-IDF value for a particular word can be calculated as follows:
  - Suppose a text review contains 100 words, wherein the word ''fine'' appears 10 times. The term frequency i.e., TF for ''fine'' is calculated as (10 / 100) = 0.1.
  - Again, let there exists 1 lakh text reviews in the corpus and the word ''fine'' appears 1000 times in whole corpus. Then, the inverse document frequency i.e., IDF is calculated as log (100,000 / 1,000) = 2.
  - Thus, the TF-IDF value is calculated as 0.1 * 2 = 0.2.

  The TF-IDF value for different words are calculated using above approach. These values are than replaced with '1' or '0' value i.e., present in the Table 3.1 while counting the CV. In TF-IDF approach the frequency of the words are considered for analysis unlike CV where only presence of the words are considered for analysis. In order to convert the text reviews to numerical vectors, combination of both the approaches are carried out in order to provide better representation of the text as compare to using the approaches separately.

3.3.3 Dataset Used

In this chapter for classification of sentiment reviews, two different movie review dataset are considered. The details of the datasets are as follows:

- **aclIMDb Dataset:** The acl Internet Movie Database (IMDb) consists 12500 positively labeled test reviews and 12500 positively labeled train reviews. Similarly there are 12500 negatively labeled test reviews and 12500 negatively labeled train reviews [126]. Apart from labeled supervised data, an unsupervised dataset is also present with 50000 reviews.

- **Polarity Dataset:** The polarity dataset consists of 1000 positive reviews and 1000 negative reviews [33]. Though the database contains both negative and positive reviews, it is not partitioned for training and testing. In order to perform the classification process, the cross validation method is being used for this dataset.

3.3.4 Data Processing Techniques

The details of processing on two datasets are explained below:

- The IMDb dataset contains separate dataset for training and testing. Thus, the training data is given as input to the MLTs for learning and based on these information, the testing dataset is being checked for its polarity i.e., either positive or negative.

- The Polarity dataset does not have a separation between training and testing data. Thus, cross validation technique is adopted for classification. Cross validation is a technique to compare algorithms by partitioning the dataset into two parts. First part, is used for learning and other part is used for validation purpose. These validation and training sets are desired to cross-over in successive rounds so that each data point needs to be validated [127]. K-fold cross validation is the basic cross validation method. In k-fold validation, dataset is partitioned into k different folds. From these k folds, k-1 folds are used for training and one fold is used for testing. 10-fold cross validation is often adopted in machine learning and classification problems, by different authors.

3.3.5 Use of Machine Learning Technique

After the transformation of the text reviews into vectors of number, they need to be processed using different machine learning techniques to obtain the classification result. In this chapter, four different MLTs are used to classify the movie reviews as discussed in Section 3.3.3. The details of these MLTs are explained as follows:

- **Naive Bayes classifier**: This method is used for both classification and training purposes [128]. This is a probabilistic classifier based on Bayes' theorem. In this chapter, three different versions of NB are used for analysis and the version of NB which provides best result is considered for comparison.

  A document is considered to be an ordered sequence of words obtained form vocabulary 'v'. The probability of a word event is independent of word context and it's

position in the document [128]. Thus, each document d_i obtained from multinomial distribution of word is independent of the length of d_i. N_{it} is the count of occurrence of w_t in document d_i. The probability of a document belonging to a class, can be obtained using the following equation:

$$P(d_i|c_j;\theta) = P(|d_i|)|d_i|! \prod_{t=1}^{|V|} \frac{P(w_t|c_j;\theta)^{N_{it}}}{N_{it}!} \tag{3.1}$$

After estimating the parameters calculated from training document, classification is performed on text document by calculating posterior probability of each class and selecting the highest number of probable classes.

The three different versions of NB are often used for classification. They are:

1. Gaussian Naive Bayes: This version of NB mainly deals with continuous data. The probability distribution for a class, $p(x = v|c)$, can be computed by plugging 'v' into an equation for a Normal distribution, parameterized by μ_c and σ_c^2, as mentioned below:
$$P(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)}{2\sigma_c^2}} \tag{3.2}$$
where
μ_c is the mean of the values in x associated with class C;
σ_c^2 is the variance of values in x associated with class C.

2. Multinomial Naive Bayes: This version of NB is often applied for text classification. The distribution is parametrized by vectors $\theta_y = (\theta_{y_1}, \ldots, \theta_{y_n})$ for each class y, where n is the number of features and θ_{y_i} is the probability $P(x_i \mid y)$ of feature 'i' appearing in a sample belonging to class y. The parameter θ_y can be estimated as follows:
$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{3.3}$$
where
$N_{yi} = \sum_{x \in T} x_i$ is the number of times feature i appears in a sample of class y in the training set T;
$N_y = \sum_{i=1}^{|T|} N_{yi}$ is the total count of all features for class y;
α is the smoothing factor;
the value of $\alpha \geq 0$.

3. Bernoulli Naive Bayes: This version of NB is used where there may be multiple features and each one is assumed to be a binary-valued variable. In text classification word occurrence vector is used for training and then for

classification. The decision rule for Bernoulli NB is as follows:

$$P(x_i|y) = P(i|y)x_i + P(1 - P(i|y))(1 - x_i) \tag{3.4}$$

The Bernoulli NB classifier explicitly penalizes the non-occurrence of a feature 'i' that is an indicator for class y, where as the multinomial variant would simply ignore a non-occurring feature.

- **Support Vector Machine Classifier (SVM):** This method analyzes data and defines decision boundaries by having hyper planes. In two category case, the hyper plane separates the document vector in one class from other class where the separation is kept as large as possible.

  For a training set with labeled pair $(x_i, y_i), i = 1, 2, ...$ where $x_i \in R^n$ and y $\in \{1, -1\}^l$, the SVM required to solve the following optimization problem may be represented as [129]:

$$\begin{aligned} &\underset{w,b,\xi}{min} \ \tfrac{1}{2}W^T W + C\sum_l^{i=1}\xi_i \\ &subject \ to \ y_i(w^T\phi(X_i) + b) \geq 1 - \xi_i, \\ &\xi_i \geq 0. \end{aligned} \tag{3.5}$$

  Here training vector x_i is mapped to higher dimensional space by ϕ. SVM requires input in the form of a vector of real numbers. Thus, the reviews of text file for classification may be converted to numeric value before it can be made applicable for SVM. After the text file is converted to numeric vector, it goes through a scaling process which manages the vectors and keep them in the range of $[1, 0]$.

  In SVM, various kernels are used for pattern analysis. There are mainly four different types of kernels used for analysis in SVM. These are as follows:

  1. Linear Kernel: The linear kernel function can be represented as follows

$$K(x_i, x_j) = x_i^T x_j. \tag{3.6}$$

     where
     x_i and x_j are the input space vector;
     x_i^T is the transpose of x_i.

  2. Polynomial Kernel: For degree 'd', the polynomial kernel function can be defined as

$$K(x_i, x_j) = \{x_i^T x_j + c\}^d \tag{3.7}$$

     where
     x_i and x_j are the input space vectors i.e., the features computed from training

sample;
'c' is a parameter used for the trade off between the highest order and lowest order polynomial.
Polynomial kernel with degree = 2, is often used in sentiment classification.

3. Gaussian Radial Basis Function (RBF) kernel: The RBF is a real valued function, whose value depends upon the distance form the origin. The RBF kernel function can be defined as follows

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||) \ \ for \ \ \gamma > 0 \tag{3.8}$$

where
x_i and x_j are the input space vector;
the value of γ can be used as $\frac{1}{2\sigma^2}$

4. Sigmoid Kernel: The sigmoid kernel function can be defined as follows

$$K(x_i, x_j) = tanh(ax_i^T x_j + b) \tag{3.9}$$

where
x_i and x_j are the input space vectors;
a>0 is the scaling parameter for the input data;
b is a shifting parameter that controls the threshold of mapping.

- **Random Forest Classifier:** This method is a combination of tree predictors. Each tree is built upon random vector values, which are sampled independently and the same application is implemented for each tree in the forest. Random forest is the collection of 'm' trees $\{T_1(x)$,...,$T_m(x)$ $\}$, where x = $\{x_1$,...,x_n $\}$ is a n dimensional vector of properties associated with each tree. The collection produces m outputs

$$\hat{Y_1} = T_1(x), \hat{Y_2} = T_2(x), ..., \hat{Y_m} = T_m(x) \tag{3.10}$$

where
$\hat{Y_m}$ is the prediction of properties of m^{th} tree.
The output obtained from all predictors are combined to obtain the final output [130].

For a collection of classifier $\{h_1(x)$,...,$h_m(x)$ $\}$ and training set obtained by random selection from random vector distribution Y and X, the marginal function can be defined as

$$mg(X, Y) = av_k I(h_k(X) = Y) - \underset{j \neq Y}{maxav_k} I(h_k(X) = j) \tag{3.11}$$

where I in the indication function. The margin suggests the range, by which the average number of votes at X,Y for correct class, betters the average votes for any

class. The bigger the margin, the better is the classification result. For a large no of trees in random forest, $h_m(X) = h(X, \Theta_m)$, where Θ_m is the independent identically distributed random vector [131].

- **Linear Discriminant Analysis Classifier (LDA):** LDA technique was first proposed by Ronald A. Fisher in 1936 which was used for two-class problem, and subsequently, the multi class LDA was proposed by C. R. Rao in 1948 [132]. This method is concerned with representing the dependent variables as linear combination of the independent variables. These linear equations are then processed to obtain the required classification result [133].

  The LDA classification steps carried out in this present approach can be explained as follows:

  i. The training and testing data are represented as matrix containing features as rows and files as column, which is represented as below:

$$train_i = \begin{bmatrix} d_{11} & d_{12} & ... \\ d_{21} & d_{22} & ... \\ ... & ... & ... \\ d_{m1} & d_{m2} & ... \end{bmatrix} \qquad test_i = \begin{bmatrix} t_{11} & t_{12} & ... \\ t_{21} & t_{22} & ... \\ ... & ... & ... \\ t_{n1} & t_{n2} & ... \end{bmatrix} \tag{3.12}$$

  The training dataset has m files and testing dataset has n files respectively. The element d_{11} represents the feature '1' in training file '1' and t_{11} represents the feature '1' in the testing file '1'.

  ii. The mean of each training and testing dataset are calculated along with the mean of the entire matrix is calculated. Let μ_{train} and μ_{test} are the mean of training and testing data respectively. μ_{total} is the mean of the entire combination and calculated as:

$$\mu_{total} = AP_{train} * \mu_{train} + AP_{test} * \mu_{test} \tag{3.13}$$

  where
  AP_{train} and AP_{test} are the apriori probability of the training data and testing data respectively.

  iii. In order to check the separation between classes, with-in class (WC) and between-class (BC) scatter are used. WC scatter can be calculated as:

$$S_{wc} = \sum_j AP_j * cov_j \tag{3.14}$$

  where cov_j can be calculated as follows

$$cov_j = (x_j - \mu_j) * (x_j - \mu_j)^T \tag{3.15}$$

BC scatter is calculated as

$$S_{bc} = \sum_{j} (\mu_j - \mu_{total}) * (\mu_j - \mu_{total})^T \tag{3.16}$$

The optimizing criteria of LDA is

$$criteria = inverse(S_{wc}) * S_{bc} \tag{3.17}$$

iv. Eigenvectors (e_1, e_2, ….., e_d) and corresponding eigenvalues (λ_1, λ_2, …., λ_d) are calculated for each scatter matrix.

v. For two class problem, two non-zero eigenvalues are generated. The eigenvectors corresponding to nonzero eigenvalues are transformed using following equation:

$$transform_j = transform^T * set_j \tag{3.18}$$

vi. After the transformation matrix is generated, the Euclidean distance is used to classify the data and can be calculated using following equation:

$$dist_n = transform_n^T * (X - \mu_{trans}) \tag{3.19}$$

where
μ_{trans} is the mean of the transformed dataset;
'n' is the class index;
'X' is the test vector.

vii. The smallest distance from the center specifies to which class the test vector belongs to.

3.3.6 Evaluation Parameters

The performance of the MLTs are often checked in order to have a comparative view. Confusion matrix also known as contingency table is helpful in visualization of performance of MLTs and shown in Table 3.1 .

Table 3.2: Confusion Matrix

	Correct Labels	
	Positive	Negative
Positive	TP (True Positive)	FP (False Positive)
Negative	FN (False Negative)	TN (True negative)

From classification point of view, the elements of confusion matrix, i.e., True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FP) values are used to

compare label of classes [134]. True Positive represents the reviews which are positive, also classified as positive by the classifier whereas False Positive are positive reviews which the classifier classifies them as negative. Similarly, True Negative represents the reviews which are negative and also classified as negative by the classifier whereas False Negative are negative reviews but classifier classifies them as positive.

Based on the data of confusion matrix, precision, recall, F-measure and accuracy are the evaluation measures used for evaluating performance of classifier.

- **Precision:** It measures the exactness of the classifier result. For binary classification problem, precision is the ratio of number of reviews correctly labeled as positive to total number of positively classified reviews where as negative predictive value (NPV) is the ratio number of examples correctly labeled as negative to total number of negative classified reviews.

$$Precision = \frac{TP}{TP + FP} \qquad NPV = \frac{TN}{TN + FN} \tag{3.20}$$

  But from the view of classification, NPV can be represented as precision for negative also.

- **Recall:** It measures the completeness of the classifier result. For binary classification problem, recall is the ratio of total number of positively labeled reviews to total reviews that are truly positive where as true negative rate (TNR) is the ratio of total number of negative labeled reviews to total reviews that are truly negative.

$$Recall = \frac{TP}{TP + FN} \qquad TNR = \frac{TN}{TN + FP} \tag{3.21}$$

  But from the view of classification, TNR can be represented as precision for negative also.

- **F-measure:** It is the harmonic mean of precision and recall. It is required to optimize the system towards either precision or recall which have a more influence on final result.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.22}$$

  Like precision and recall for both negative and positive reviews, similarly the F-measure also obtained for both positive and negative values.

- **Accuracy:** It is the most common measure of classification accuracy. It can be calculated as the ratio of correctly classified reviews to total number of reviews.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.23}$$

3.4 Proposed Approach

In this chapter, supervised machine learning techniques are applied on two different review datasets such as IMDb and Polarity. The datasets are then preprocessed and transformed into numerical vectors. These vectors are then processed by different MLTs and finally the performance of the MLTS are assessed using different parameters. The stepwise detailed elaboration of the proposed approach is shown in Figure 3.1.

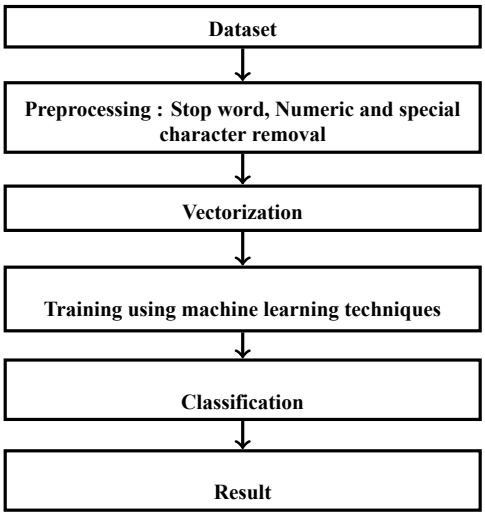

Figure 3.1: *Diagrammatic view of the proposed approach*

Step 1. The two different datasets considered for analysis are:

- aclIMDb dataset: It consists of 12500 positive and 12500 negative review for training and also 12500 positive and 12500 negative review for testing [32].
- Polarity dataset: it consists of 1000 positive and 1000 negative review for analysis [33].

Step 2. The text reviews in the dataset consist of absurd information which need to be removed from the reviews before it is considered for classification. The absurd information are :

- Stop words: Stop words do not play any role in determining the sentiment. The list of English stop words are collected from site ‘‘http://norm.al/2009/04/14/list-of-english-stop-words/’’. These stop words are searched in the text and if found out removed from the text as they have no role in the sentiment analysis of text.
- Numeric and special character: In the text reviews, there are different numeric values like 1, 2, 3, ... etc. and special characters such as , #, $,% etc., which do not have any effect on the analysis, but they create confusion while converting text file to numeric vector.

- HTML tags, http:// https:// and Email IDs: These types of words used in text need to be removed as they do not help in sentiment analysis and also create confusion while analyzing the text.
- Lowering the case: It is sometimes observed that some of the words in text present in the reviews do not contain words or characters in uniform case. Thus, these words or reviews need to be converted into a uniform case for the ease in processing of the text. In this case, all the texts are converted into lower case to maintain uniformity.
- Stemming: It is a process of obtaining the root word from any word. For example: the words such as plays, playing, played all have the word play as its root. So, while instead of analyzing all above words, the word play can be used. Thus, the root word for each word is being obtained and based on the root word only, the classification is done. The PorterStemmer tool is used for the stemming purpose [45].

Step 3. After the preprocessing of text reviews, the text reviews are converted to numeric vectors. The methods used for the conversion of text file to numeric vectors are as follows:

- CV: It converts the text reviews into a matrix of token counts. It implements both tokenization and occurrence counting.
- TF-IDF: It suggests the importance of the word to the document and to the whole corpus. Term frequency informs about the frequency of a word in a document and IDF informs about the frequency of the particular word in whole corpus.

Step 4. After the text reviews are transformed into numeric vectors, these are considered as input to four different supervised machine learning algorithms for classification purpose. The algorithms are as follows:

- NB: Using probabilistic classifier and pattern learning, it examines the set of documents and classifies accordingly [128].
- SVM: SVM analyzes data and defines decision boundaries by having hyper-planes. In two category case, the hyper-plane separates the document vector in one class from other class where the separation is kept as large as possible [129].
- RF: Random forest constitutes a set of multiple decision trees for each input vector at training time. The tree votes for the correct class, and the larger the number of votes obtained, better is the classification result [130].

- LDA: In this method the dependent variables are represented as a linear combination of the independent variables. These linear equations are then solved to obtain the required classification result [133].

Step 5. Results obtained on two different datasets are furnished below:

- aclIMDb dataset: This dataset has a separate dataset from training and testing purposes. The MLTs get trained using the training dataset and based on the information obtained from training, the testing dataset is tested. Different evaluation parameters are used to evaluate performance of MLTs.

  NB classifier: As discussed in Section 3.3.5, three different versions of NB classifiers are considered for analysis. The confusion matrix and other performance parameters obtained after analysis for each cases are shown in Table 3.3

Table 3.3: Confusion matrix, Evaluation Parameters and Accuracy for Naive Bayes Classifier

Method Used	Confusion Matrix			Evaluation Parameters			Accuracy
Gaussian Naïve Bayes		Correct Label		Precision	Recall	F-measure	0.757
		Positive	Negative				
	Positive	8259	3744	0.74	0.89	0.81	
	Negative	1320	10680	0.86	0.77	0.81	
Multinomial Naïve Bayes		Correct Label		Precision	Recall	F-measure	**0.831**
		Positive	Negative				
	Positive	11107	1393	0.87	0.77	0.82	
	Negative	2834	9666	0.8	0.89	0.84	
Bernoulli Naïve Bayes		Correct Label		Precision	Recall	F-measure	0.827
		Positive	Negative				
	Positive	11049	1451	0.87	0.77	0.82	
	Negative	2870	9630	0.79	0.88	0.84	

RF classifier: The confusion matrix and other performance evaluation parameters obtained after analysis of the reviews are shown in Table 3.4

Table 3.4: Confusion matrix, Evaluation Parameters and Accuracy for Random Forest Classifier

Method Used	Confusion Matrix			Evaluation Parameters			Accuracy
Random Forest		Correct Label		Precision	Recall	F-measure	**0.88884**
		Positive	Negative				
	Positive	11161	1339	0.89	0.88	0.88	
	Negative	1440	11060	0.88	0.89	0.89	

SVM classifier: As discussed in Section 3.3.5, four different kernels of SVM are considered for analysis. The confusion matrix and other performance parameters obtained after analysis for each cases are shown in Table 3.5

Table 3.5: Confusion matrix, Evaluation Parameters and Accuracy for Support Vector Machine Classifier

Method Used	Confusion Matrix			Evaluation Parameters			Accuracy
Linear Kernel		Correct Label		Precision	Recall	F-measure	**0.8842**
		Positive	Negative				
	Positive	11018	1482	0.88	0.89	0.88	
	Negative	1413	11087	0.89	0.88	0.88	
Polynomial Kernel		Correct Label		Precision	Recall	F-measure	0.8294
		Positive	Negative				
	Positive	10304	2196	0.83	0.83	0.83	
	Negative	2067	10433	0.83	0.82	0.83	
Gaussian RBF kernel		Correct Label		Precision	Recall	F-measure	0.8382
		Positive	Negative				
	Positive	11123	1377	0.88	0.79	0.83	
	Negative	2666	9834	0.81	0.89	0.85	
Sigmoid Kernel		Correct Label		Precision	Recall	F-measure	0.862
		Positive	Negative				
	Positive	11088	1412	0.88	0.84	0.86	
	Negative	2030	10470	0.85	0.89	0.87	

LDA classifier: The confusion matrix and other performance evaluation parameters obtained after analysis of the reviews are shown in Table 3.6

Table 3.6: Confusion matrix, Evaluation Parameters and Accuracy for Linear Discriminant Analysis Classifier

Method Used	Confusion Matrix			Evaluation Parameters			Accuracy
Linear Discriminant Analysis		Correct Label		Precision	Recall	F-measure	**0.86976**
		Positive	Negative				
	Positive	10993	1507	0.88	0.86	0.87	
	Negative	1749	10751	0.86	0.89	0.87	

The following Figure 3.2 provides a comparison between the accuracy values obtained by different MLTs using aclIMDb dataset.

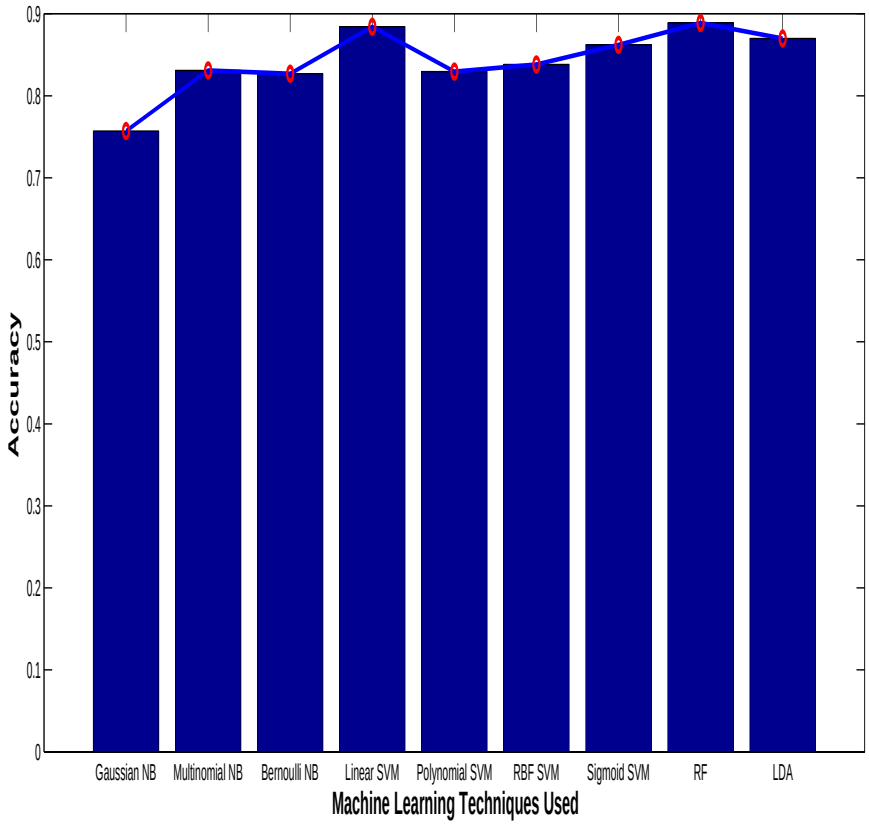

Figure 3.2: *Comparison of Accuracy values of Proposed MLTs using IMDb dataset*

From the Figure 3.2, it can be analyzed that the RF classifier shows the best result among all approaches.

- Polarity Dataset: Unlike IMDb dataset, where there is separation between the training and testing data, polarity dataset has positive reviewed dataset and negative reviewed dataset. Thus, K fold cross-validation technique is adopted for classification. In this case, ten fold cross validation is used for classification of reviews of polarity dataset. The Table 3.7 shows the values of average accuracy after each fold using the different MLTs.

Table 3.7: Classification accuracy obtained after 10 fold cross validation on Polarity dataset

Classification Algorithm Used		Average Accuracy
Naive Bayes Classifier	Gaussian NB	0.795
	Multinomial NB	0.895
	Bernoulli NB	0.855
Support Vector Machine Classifier	Linear Kernel	0.940
	Polynomial Kernel	0.910
	Gaussian RBF Kernel	0.900
	Sigmoid Kernel	0.903
Random Forest Classifier		0.950
Linear Discriminant Analysis Classifier		0.920

The following Figure 3.3 provides a comparison between the accuracy values obtained by different MLTs using polarity dataset.

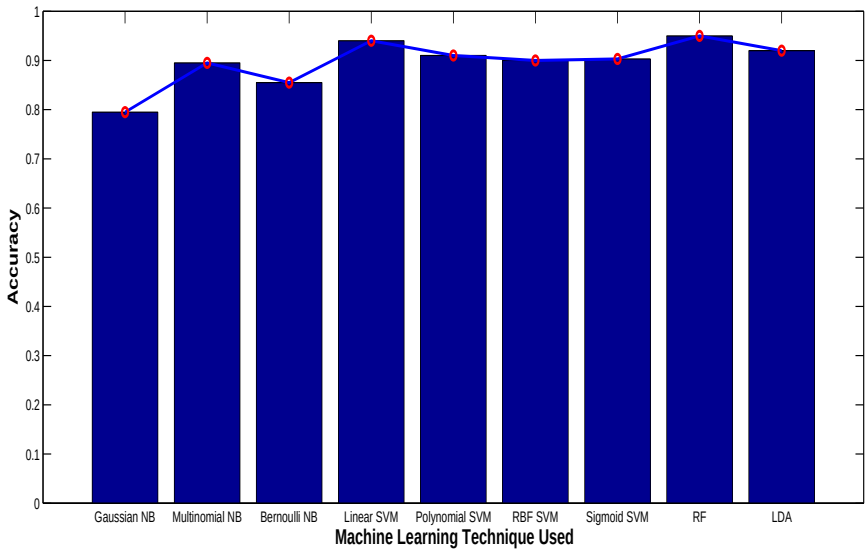

Figure 3.3: *Comparison of Accuracy values of Proposed MLTs using polarity dataset*

3.5 Performance Evaluation

Table 3.8 and Figure 3.4 show the comparison of accuracy values using the proposed approach with other approaches as available in literature using IMDb dataset. From the table, it can be observed that the approach adopted in this chapter i.e., the combination of both countvectorizer and TF-IDF for transformation of input text into numeric value yields better result in comparison with results obtained by authors of different articles in literature. It is further found out that values of accuracy obtained using Naive Bayes, Support Vector Machine, Random Forest and LDA algorithm are 0.831, 0.884, 0.888 and 0.869 respectively.

Table 3.8: Comparative results obtained among different literature using IMDb Dataset

Classifier Used	**Classification Accuracy**					
	Pang et al. [8]	**Salvetti et al.** [34]	**Mullen and Collier** [36]	**Beineke et al.** [35]	**Matsumoto et al.** [56]	**Proposed Approach**
Naive Bayes	0.815	0.796	⊗	0.659	⊗	0.831
Support Vector Machine	0.659	⊗	0.86	⊗	0.883	0.884
Random Forest	⊗	⊗	⊗	⊗	⊗	**0.888**
Linear Discriminant Analysis	⊗	⊗	⊗	⊗	⊗	0.869

'⊗' mark indicates that the technique not considered by the author in their respective paper

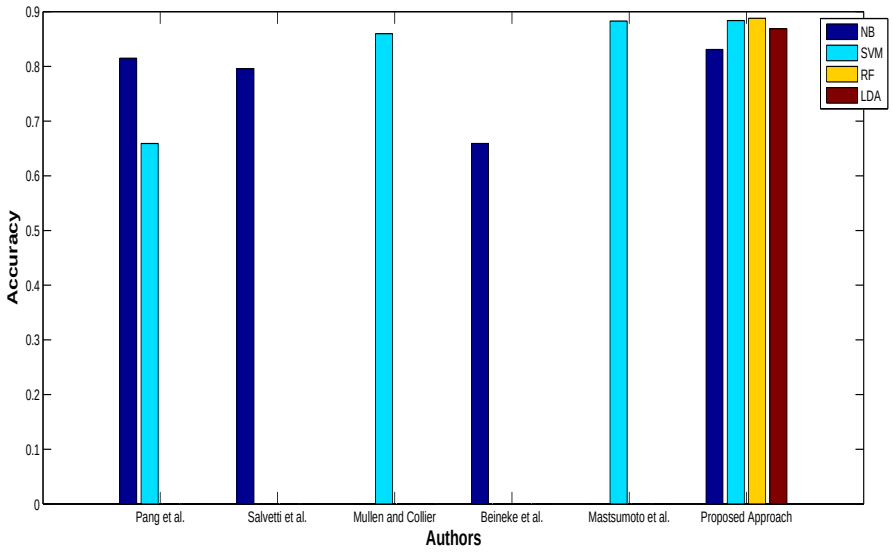

Figure 3.4: *Comparison of Accuracy of different literatures using IMDb dataset*

Table 3.9 and Figure 3.5 show the comparison of the proposed approach with other approaches followed by different author of literature using Polarity dataset. In order to

Table 3.9: Comparative result obtained among different literature using Polarity dataset

Classifier Used	Classifier Accuracy						
	Pang and Lee [33]	Matsumoto et al. [56]	Aue and Gamon [135]	Read [136]	Kennedy and Inkpen [137]	Whitelaw et al. [138]	Proposed Approach
Naive Bayes	0.864	⊗	⊗	0.789	⊗	⊗	0.895
Support Vector Machine	0.872	0.937	0.905	0.815	0.862	0.902	0.940
Random Forest	⊗	⊗	⊗	⊗	⊗	⊗	0.950
Linear Discriminant Analysis	⊗	⊗	⊗	⊗	⊗	⊗	0.920

'⊗' mark indicates that the technique not considered by the author in their respective paper

classify the review, the method of cross validation is used here. Apart from the work done by author i.e., Aue and Gamon [135], a good number of authors have used 10 fold cross validation for classification purpose. Aue and Gamon have used 5 fold cross validation for classification. In 10 fold cross validation 90% for the reviews are considered for training and rest 10 % are used for testing. Similar to the case of IMDb dataset, the accuracy result obtained in polarity dataset is comparably better than other methods as shown in Table 3.9.

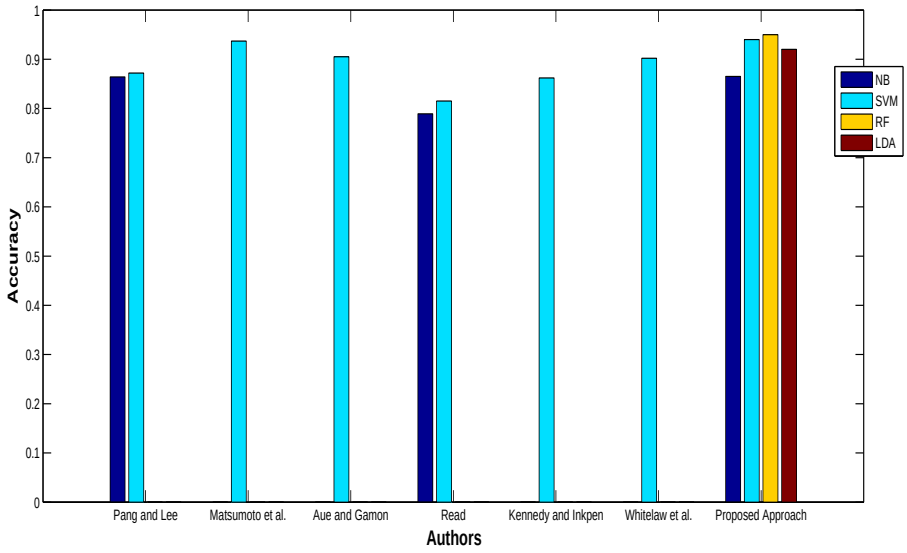

Figure 3.5: *Comparison of Accuracy of different literatures using Polarity dataset*

From Table 3.8 and Table 3.9, it is evident that Random Forest classifier shows better result in comparison with other classifiers. The reasons may be attributed as follows:

- It runs efficiently on large data bases.
- It generates an internal unbiased estimate of the generalization error as the forest building progresses.

- It provides methods for balancing error in unbalanced data sets.
- It effectively estimates missing data.
- Prototypes are computed which give information about the relation between the variables and the classification.

3.6 Summary

In this chapter, two movie review datasets i.e., IMDb and polarity are considered for analysis. The two datasets are considered as IMDb has separate data for training and testing while polarity dataset has no separation between training and testing data. Thus, 10 fold cross validation technique has been considered for its analysis. Four different MLTs i.e., NB with three variants, SVM with four different kernels, Random Forest and LDA are used in this chapter. The time complexity of these methods are shown as follows:

- Naïve Bayes: O(n * m)
- SVM: O(n * m)
- Random Forest (RF): O(m * n logn)
- LDA: O(m * n * logn)

where 'n' is the number of reviews present and 'm' is no of classes, in this case the value of 'm' is two as positive and negative classes are considered. Four different MLTs are used as NB uses probabilistic Bayesian method for classification, SVM uses kernel based system for classification, RF uses ensemble method, finally LDA uses discriminant analysis approach for classification Among these approaches, RF shows the best result in both the datasets.

Chapter 4

Classification of Sentiment Reviews using N-gram Machine Learning Approach

In this chapter, the sentiment classification of the movie reviews are improved by adding the n-gram feature to the classification. The application of the proposed approach on a standard dataset is discussed along with the accuracy results obtained using different MLTs and the result of the proposed approach is compared with existing result.
The rest of the chapter is organized as follows:
Section 4.1 provides a brief introduction to the proposed approach and the contribution of the chapter. Section 4.2 Indicates the motivation for the proposed approach. Section 4.3 discusses about different techniques to transform text data into numerical vectors, classification techniques, and performance evaluation parameters. Section 4.4 highlights the proposed approach. Section 4.5 informs about the output obtained after implementation of the proposed approaches. Section 4.6 compares the performance of the proposed approach with present literatures. Finally, Section 4.7 summarizes the Chapter.

4.1 Introduction

Sentiment analysis is concerned with analysis of text reviews about any product and helps to provide any meaningful information to others. During the process of analysis, each word is considered as a feature. In Chapter 3 each word is considered as a single unit of analysis but in this chapter an attempt is made to combine more than one word i.e., two words (bigram) and three words (trigram) for the analysis of the reviews. The contribution of the chapter can be stated as follows:

i. Different machine learning algorithms are proposed for the classification of movie reviews of IMDb dataset [126] using n-gram techniques viz., Unigram, Bigram, Trigram, combination of unigram and bigram, bigram and trigram, and unigram and bigram and trigram .

ii. Four different machine learning techniques such as Naive Bayes, Maximum Entropy, Support Vector Machine, and Stochastic Gradient Descent are used for classification purpose using the n-gram approach.

iii. The performance of the machine leaning techniques are evaluated using parameters like precision, recall, f-measure, and accuracy. The results obtained in this chapter indicate, the higher values of accuracy when compared with studies made by other authors.

4.2 Motivation for the proposed approach

The Section 2.3 discusses about Sentiment Classification using n-gram MLTs and the Table 2.2 provides a comparative analysis of those papers. These information help to identify some possible research areas which can be extended further. The following aspects have been considered for carrying out further research.

i. A good number of authors apart from Pang et al. [8], and Matsumoto et al. [56], have considered unigram approach to classify the reviews. This approach provides comparatively better result, but in some cases it does not yield suitable result. The comment ‘‘ The item is not good ’’, when analyzed using unigram approach, provides the polarity of sentence as neutral with the presence of one positive polarity word ‘good’ and one negative polarity word ‘not’. But when the statement is analyzed using bigram approach, it gives the polarity of sentence as negative due to the presence of words ‘‘not good’’, which is correct. So, when a higher level of n-gram is considered, the result is expected to be better. Thus, analyzing the research outcome of several authors, this study makes an attempt to extend the sentiment classification using unigram, bigram, trigram, and their combinations for classification of movie reviews.

ii. Also a number of authors have used part-of-speech (POS) tags for classification purpose. But it is observed that the POS tag for a word is not fixed and it changes as per the context of their use. For example, the word ‘book’ can have the POS ‘noun’ when used as reading material where as in case of ‘‘ticket booking’’ the POS is verb. Thus, in order to avoid confusion, instead of using POS as a parameter for classification, the word as a whole may be considered for classification.

iii. Most of the machine learning algorithms work on the data represented as matrix of numbers. But the sentiment data are always in text format. So, it needs to be converted to number matrix. Different authors have considered TF or TF-IDF to convert the text into matrix on numbers. But in this chapter, in order to convert the text data into matrix of numbers, the combination of TF-IDF and CountVectorizer has been applied. The rows of the matrix of numbers represents a particular text file where as its column represents each word / feature present in that respective file which is shown in Table 3.2.

4.3 Methodology Adopted

This section discusses about the different methodologies adopted for classification of sentiment reviews.

4.3.1 Types of sentiment classification

According to the Section 3.3.1, there exist two types of classification technique i.e., binary and multi-class sentiment analysis. The most used technique is binary classification which is adopted in this chapter for classification of movie reviews.

4.3.2 Transformation of Text Data into Numerical values / matrix

The sentiment reviews are mainly in the text format and the MLTs need the data in the form of numerical vectors only for classification. The Section 3.3.2 discusses about the different transformation techniques for converting the text reviews into numerical vectors. The two approaches for transformation i.e., countvectorizer and TF-IDF are used simultaneously in this chapter for better representation of the text as numerical vector.

4.3.3 Dataset used

In this chapter, for the sentiment classification of movie reviews, aclIMDb dataset is used. This dataset contain separate data for training and testing. Both training and testing dataset contain 12500 positive and 12500 negative reviews respectively. Apart from them, it also contain 50000 reviews which are unlabeled.

As there is a separation between the training and testing data, the training data is used for training the machine learning technique and based on the information obtained from training, the MLTs test the testing data.

4.3.4 Machine Learning Techniques Used

After the transformation of text reviews to numerical vectors, those are as given as input to the MLTs for classification purpose. In this chapter, four different MLTs are used for classification. These techniques are Naive Bayes, Support Vector Machine, Maximum Entropy (ME), and Stochastic Gradient Descent (SGD).

- Naive Bayes Classifier: As discussed in Section 3.3.5, NB is being used for both classification and training purposes. This is a probabilistic classifier based on Bayes' theorem. In Chapter 3, three different versions of NB is used. But in this chapter, only the multinomial NB is used, as this approach has shown a better classification accuracy value in comparison with those of other versions of NB.

- Support Vector Machine Classifier: As discussed in Section 3.3.5, SVM analyzes data and defines decision boundaries by having hyper planes. In two category case, the hyper plane separates the document vector in one class from other class where the separation is kept as large as possible. In Chapter 3, four different kernels are proposed and among them linear kernel is observed to yield best result. Thus, in this chapter for classification point of view, linear kernel based SVM is only used on IMDb dataset.

- **Maximum Entropy (ME) method:** In this method, the training data is used to set constraint on conditional distribution [139]. Each constraint is used to express characteristics of training data. Maximum Entropy (ME) value in terms of exponential function can be expressed as:

$$P_{ME}(c|d) = \frac{1}{Z(d)} exp(\sum_{i} \lambda_{i,c} f_{i,c}(d,c)) \tag{4.1}$$

where
$P_{ME}(c|d)$ refers to probability of document 'd' belonging to class 'c';
$f_{i,c}(d,c)$ is the feature / class function for feature f_i and class c;
$\lambda_{i,c}$ is the parameter to be estimated;
$Z(d)$ is the normalizing factor.

In order to use ME, a set of features is needed to be selected. For text classification purpose, word counts are considered as features. Feature / class function may be instantiated as follows:

$$f_{i,c'}(d,c) = \begin{cases} 0 & if c \neq c' \\ \dfrac{N(d,i)}{N(d)} & otherwise \end{cases} \tag{4.2}$$

where
$f_{i,c'}(d,c)$ refers to features in word-class combination in class 'c' and document 'd';
$N(d,i)$ represents the occurrence of feature 'i' in document 'd';
$N(d)$ number of words in 'd'. As per the expression, if a word occurs frequently in a class, the weight of word-class pair becomes higher in comparison with other pairs. These highest frequency word-class pairs are considered for classification purpose.

- **Stochastic Gradient Descent (SGD) method:** This method is used when the training data size is observed to be large. In SGD method instead of computing the gradient, each iteration estimates the value of gradient on the basis of single randomly picked example as considered by Léon Bottou [140].

$$w_{t+1} = w_t - \gamma_t \nabla_w Q(z_t, w_t) \tag{4.3}$$

The stochastic process $\{w_t, t = 1, 2, ...\}$ depends on randomly picked example at each iteration, where $Q(z_t, w_t)$ is used to minimize the risk and γ_t is the learning rate. The convergence of SGD gets effected by the noisy approximation of the gradient. If learning rate decreases slowly, the parameter estimate w_t decreases equally slowly; but if rate decreases too quickly, the parameter estimate w_t takes significant amount of time to reach the optimum point.

- **N-gram model:** It is a method of checking 'n' continuous words or sounds from a given sequence of text or speech. This model helps to predict the next item in a sequence. In sentiment analysis, the n-gram model helps to analyze the sentiment of the text or document. Unigram refers to n-gram of size 1, Bigram refers to n-gram of size 2, Trigram refers to n-gram of size 3. Higher n-gram refers to four-gram, five-gram, and so on. The n-gram method can be explained using following example:

  A typical example of a sentence may be considered as ''The movie is not a good one''.

  - Its unigram: '''The','movie','is', 'not', 'a', 'good','one''' where a single word is considered.
  - Its bigram: '''The movie','movie is', 'is not', 'not a', 'a good', 'good one' '' where a pair of words are considered.
  - Its trigram: '' 'The movie is', 'movie is not', 'is not a', 'not a good','a good one' '' where a set of words having count equal to three is considered.

4.3.5 Parameters used for Performance Evaluation

As discussed in Section 3.3.6, confusion matrix is used for evaluation of the performance of the MLTs. The Table 3.1 shows the confusion matrix, which include terms like true positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Using this terms, some parameters like precision, NPV, recall, TNR, F-measure and accuracy are calculated to evaluate the performance of the MLTs.

4.4 Proposed Approach

The movie reviews of IMDb dataset is processed to remove the stop words and unwanted information. The text data is then transformed to numerical vector using vectorization techniques. Further, training of the dataset is carried out using MLTs and based on that testing is done using n-gram approach. The stepwise detailed elaboration of the proposed approach is shown in Figure 4.1 .

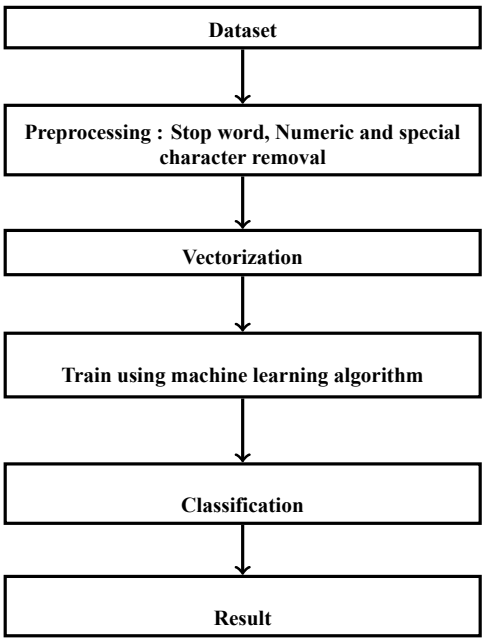

Figure 4.1: *Diagrammatic view of the proposed approach*

The detailed description of the steps are mentioned below:

Step 1: The aclIMDb dataset consisting of 12,500 positive and 12,500 negative reviews for training and also 12,500 positive and 12,500 negative reviews for testing [126], is taken into consideration.

Step 2: The text reviews sometimes consist of absurd data, which need to be removed, before considered for classification. The identified absurd data are:

- Stop words: They do not play any role in determining the sentiment.
- Numeric and special character: In the text reviews, it is often observed that there are different numeric (1,2,...5 etc.) and special characters (@, #, $,% etc.) present, which do not have any effect on the analysis. But they often create confusion during conversion of text file to numeric vector.
- HTML tags, http:// https:// and Email IDs: These information need to be removed as they do not help in sentiment analysis and also create confusion while analyzing the text.
- Lowering the case: The text present in the reviews does not contain words or characters in uniform case. Thus, these words or reviews need to be converted into a uniform case for the easy processing of the text. In this case, all the texts are converted into lower case to maintain uniformity.
- Stemming: It is a process of obtaining the root word from a word. A particular word can be used in many forms, i.e., verb, adjective, and noun with a little change in the root word. Thus, these words need to be transformed into root

words which checks the multiple entries into the list of features / words and also reduce the load on classifier.

Step 3: After the preprocessing of text reviews, they need to be converted to a matrix of numeric vectors. The following methodologies are considered for conversion of text file to numeric vectors:

- CountVectorizer: It converts the text reviews into a matrix of token counts. It implements both tokenization and occurrence counting. The output matrix obtained after this process is a sparse matrix.
- TF-IDF: It suggests the importance of the word to the document and whole corpus. Term frequency informs about the frequency of a word in a document and IDF informs about the frequency of the particular word in whole corpus [125].

In this chapter, for the transformation of the text reviews into numerical vectors both methods are used simultaneously.

Step 4: After the text reviews are converted to matrix of numbers, these matrices are considered as input for the following four different supervised machine learning algorithms for classification purpose.

- Naive Bayes (NB) method: Using probabilistic classifier and pattern learning, the set of documents are classified [128].
- Maximum Entropy (ME) method: The training data are used to set constraint on conditional distribution [139]. Each constraint is used to express characteristics of training the data. These constraints then are used for testing the data.
- Stochastic Gradient Descent (SGD) method: SGD method is used when the training data size is mostly large in nature. Each iteration estimates the gradient on the basis of single randomly picked example [140].
- Support Vector Machine (SVM) method: Data are analyzed and decision boundaries are defined by having hyper planes. In two category case, the hyper plane separates the document vector of one class from other classes, where the separation is maintained to be large as possible [129].

Step 5: As mentioned in step 1, the movie reviews of aclIMDb dataset is considered for analysis, using the machine learning algorithms as discussed in step 4. Then different variation of the n-gram methods i.e., unigram, bigram, trigram, unigram + bigram, unigram + trigram, and unigram + bigram + trigram have been implemented to obtain the result as shown in Section 4.5.

Step 6: The results obtained from this analysis are compared with the results available in other literatures and are shown in Section 4.6.

4.5 Implementation

- **Application of Naive Bayes method:** The confusion matrix and various evaluation parameters such as precision, recall, f-measure, and accuracy values obtained after classification using NB n-gram techniques are shown in Table 4.1.

Table 4.1: Confusion Matrix, Evaluation Parameter and Accuracy for Naive Bayes n-gram classifier

Method	**Confusion Matrix**			**Evaluation Parameter**			**Accuracy**
Unigram		Correct Labels		Precision	Recall	F-Measure	83.652
		Positive	Negative				
	Positive	11025	1475	0.88	0.81	0.84	
	Negative	2612	9888	0.79	0.87	0.83	
Bigram		Correct Labels		Precision	Recall	F-Measure	84.064
		Positive	Negative				
	Positive	11156	1344	0.89	0.81	0.85	
	Negative	2640	9860	0.79	0.88	0.83	
Trigram		Correct Labels		Precision	Recall	F-Measure	70.532
		Positive	Negative				
	Positive	10156	2344	0.81	0.67	0.73	
	Negative	5023	7477	0.6	0.76	0.67	
Unigram + Bigram		Correct Labels		Precision	Recall	F-Measure	86.004
		Positive	Negative				
	Positive	11114	1386	0.89	0.84	0.85	
	Negative	2113	10387	0.83	0.88	0.85	
Bigram + Trigram		Correct Labels		Precision	Recall	F-Measure	83.828
		Positive	Negative				
	Positive	11123	1377	0.89	0.81	0.85	
	Negative	2666	9834	0.79	0.88	0.83	
Unigram + Bigram + Trigram		Correct Labels		Precision	Recall	F-Measure	86.232
		Positive	Negative				
	Positive	11088	1412	0.89	0.85	0.87	
	Negative	2030	10470	0.84	0.88	0.86	

The following Figure 4.2 shows a comparative analysis of accuracy obtained using different Naive Bayes based n-gram techniques.

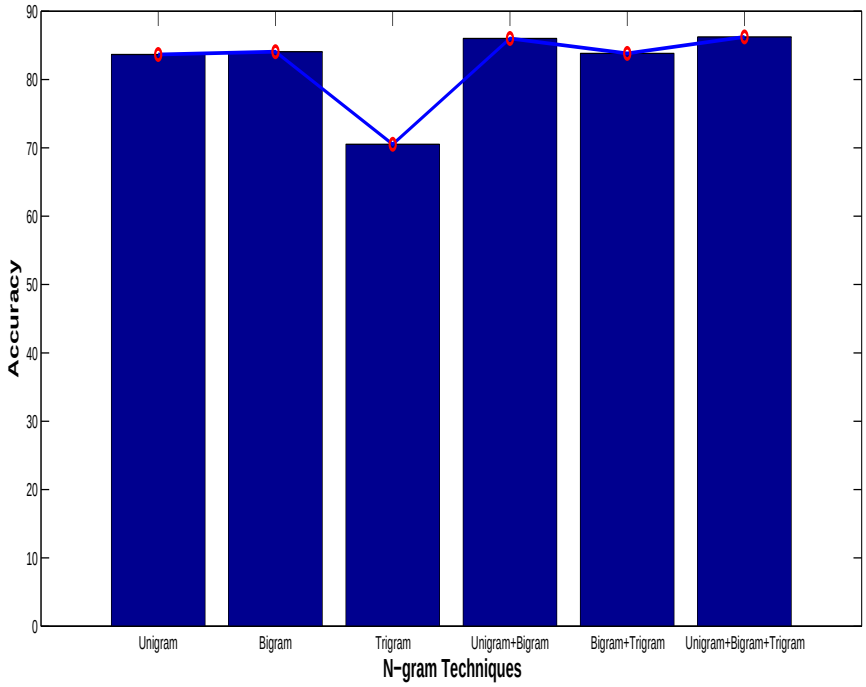

Figure 4.2: *Comparison of Accuracy values of Naive Bayes N-gram classifier*

From Table 4.1 and Figure 4.2, it can be analyzed that the accuracy value obtained using bigram is better than value obtained using techniques such as unigram and trigram. Naive Bayes method is a probabilistic method, where the features are independent of each other. Hence, when analysis is carried out using ‘‘single word (unigram)’’ and ‘‘double word (bigram)’’, the accuracy value obtained is comparatively better than that obtained using trigram. But when ‘triple word (trigram)’ is being considered for analysis of features, words are repeated a number of times; thus, it affects the probability of the document. For example: for the statement ‘‘it is not a bad movie”, the trigram ‘‘it is not’’, and ‘‘is not a’’ show negative polarity, where as the sentence represents positive sentiment. Thus, the accuracy of classification decreases. Again, when the trigram model is combined with unigram or bigram or unigram + bigram, the impact of trigram makes the accuracy value comparatively low.

- **Application of Maximum Entropy method:** The confusion matrix and evaluation parameters such as precision, recall, f-measure, and accuracy values obtained after classification using ME n-gram techniques are shown in Table 4.2.

Table 4.2: Confusion Matrix, Evaluation Parameter and Accuracy for Maximum Entropy n-gram classifier

Method	**Confusion Matrix**			**Evaluation Parameter**			**Accuracy**
Unigram		Correct Labels		Precision	Recall	F-Measure	88.48
		Positive	Negative				
	Positive	11011	1489	0.88	0.89	0.88	
	Negative	1391	11109	0.89	0.88	0.88	
Bigram		Correct Labels		Precision	Recall	F-Measure	83.228
		Positive	Negative				
	Positive	10330	2170	0.83	0.84	0.83	
	Negative	2023	10477	0.84	0.83	0.83	
Trigram		Correct Labels		Precision	Recall	F-Measure	71.38
		Positive	Negative				
	Positive	8404	4096	0.67	0.73	0.70	
	Negative	3059	9441	0.76	0.70	0.73	
Unigram + Bigram		Correct Labels		Precision	Recall	F-Measure	88.42
		Positive	Negative				
	Positive	11018	1482	0.88	0.89	0.88	
	Negative	1413	11087	0.89	0.88	0.88	
Bigram + Trigram		Correct Labels		Precision	Recall	F-Measure	82.948
		Positive	Negative				
	Positive	10304	2196	0.82	0.83	0.83	
	Negative	2067	10433	0.83	0.83	0.83	
Unigram + Bigram + Trigram		Correct Labels		Precision	Recall	F-Measure	83.36
		Positive	Negative				
	Positive	11006	1494	0.88	0.89	0.88	
	Negative	2666	9834	0.78	0.87	0.82	

The following Figure 4.3 shows a comparative analysis of accuracy obtained using different Naive Bayes based n-gram techniques.

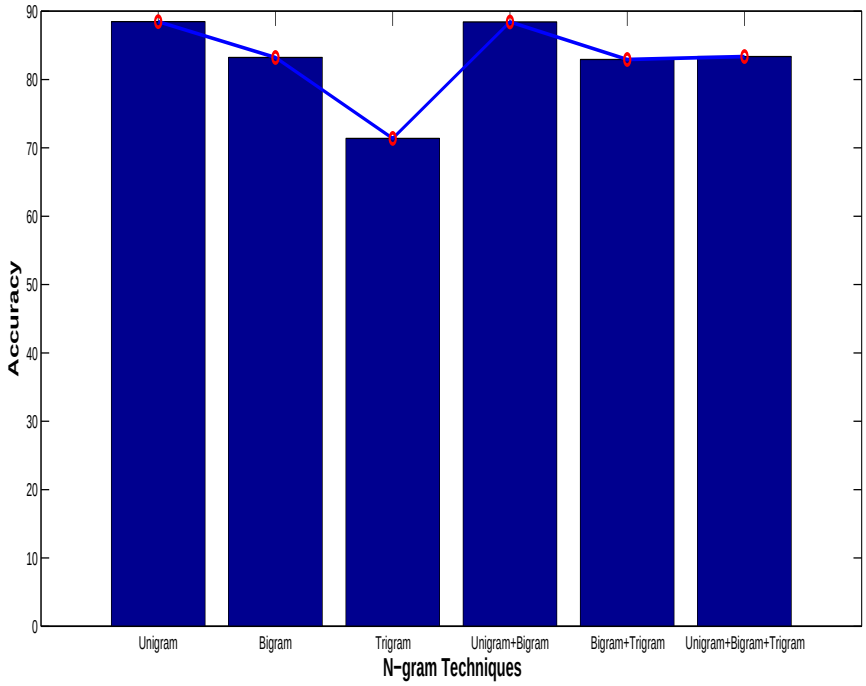

Figure 4.3: *Comparison of Accuracy values of different n-gram technique using ME*

As represented in the Table 4.2 and Figure 4.3, it may be analyzed that the accuracy value obtained using unigram is better than the values obtained using bigram and trigram. As ME algorithm based on conditional distribution and word-class pair help to classify the review, unigram method which considers single word for analysis, provides best result in comparison with other methods. In both bigram and trigram methods, the negative or positive polarity word appears more than once; thus, affecting the classification result. The bigram and trigram methods when combined with unigram and between themselves, the accuracy values of various combinations are observed to be low.

- **Application of Support Vector Machine method:** The confusion matrix and evaluation parameters such as precision, recall, f-measure, and accuracy values obtained after classification using SVM n-gram techniques are shown in Table 4.3.

Table 4.3: Confusion Matrix, Evaluation Parameter and Accuracy for Support Vector Machine n-gram classifier

Method	**Confusion Matrix**			**Evaluation Parameter**			**Accuracy**
Unigram		Correct Labels		Precision	Recall	F-Measure	86.976
		Positive	Negative				
	Positive	10993	1507	0.88	0.86	0.87	
	Negative	1749	10751	0.86	0.88	0.87	
Bigram		Correct Labels		Precision	Recall	F-Measure	83.872
		Positive	Negative				
	Positive	10584	1916	0.85	0.83	0.84	
	Negative	2116	10384	0.83	0.84	0.84	
Trigram		Correct Labels		Precision	Recall	F-Measure	70.204
		Positive	Negative				
	Positive	8410	4090	0.67	0.71	0.69	
	Negative	3359	9141	0.73	0.69	0.71	
Unigram + Bigram		Correct Labels		Precision	Recall	F-Measure	88.884
		Positive	Negative				
	Positive	11161	1339	0.89	0.89	0.89	
	Negative	1440	11060	0.88	0.89	0.89	
Bigram + Trigram		Correct Labels		Precision	Recall	F-Measure	83.636
		Positive	Negative				
	Positive	10548	1952	0.84	0.83	0.84	
	Negative	2139	10361	0.83	0.84	0.84	
Unigram + Bigram + Trigram		Correct Labels		Precision	Recall	F-Measure	88.944
		Positive	Negative				
	Positive	11159	1341	0.89	0.89	0.89	
	Negative	1423	11077	0.89	0.89	0.89	

The following Figure 4.4 shows a comparative analysis of accuracy obtained using different Support Vector Machine based n-gram techniques.

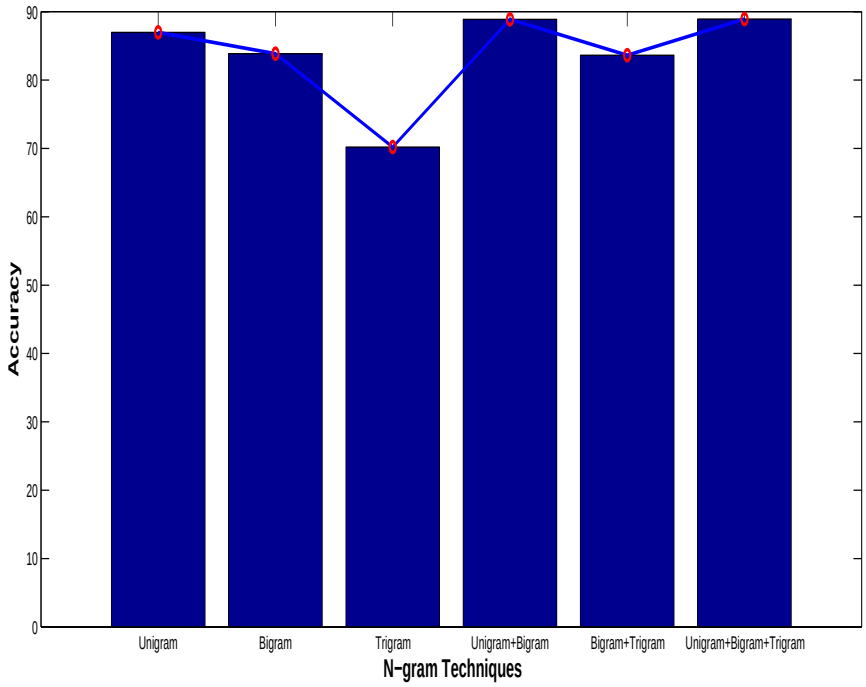

Figure 4.4: *Comparison of Accuracy values of different n-gram technique using SVM*

As exhibited in Table 4.3 and Figure 4.4, it may be analyzed that the accuracy value obtained using unigram is better than the value obtained using bigram and trigram. As SVM method is a non-probabilistic linear classifier and trains model to find hyperplane in order to separate the dataset, the unigram model which analyzes single words for analysis gives better result. In bigram and trigram, there exists multiple word combinations, which, when plotted in a particular hyperplane, confuses the classifier and thus, it provides a less accurate result in comparison with the value obtained using unigram. Thus, the less accurate bigram and trigram, when combined with unigram and with each other also, provide a less accurate result.

- **Application of Stochastic Gradient Descent method:** The confusion matrix and evaluation parameters such as precision, recall, f-measure, and accuracy values obtained after classification using SGD n-gram techniques are shown in Table 4.4.

Table 4.4: Confusion Matrix, Evaluation Parameter and Accuracy for Stochastic Gradient Descent n-gram classifier

Method	**Confusion Matrix**			**Evaluation Parameter**			**Accuracy**
Unigram		Correct Labels		Precision	Recall	F-Measure	85.116
		Positive	Negative				
	Positive	9860	2640	0.79	0.90	0.84	
	Negative	1081	11419	0.91	0.81	0.86	
Bigram		Correct Labels		Precision	Recall	F-Measure	95
		Positive	Negative				
	Positive	12331	169	0.99	0.92	0.95	
	Negative	1081	11419	0.91	0.99	0.95	
Trigram		Correct Labels		Precision	Recall	F-Measure	58.408
		Positive	Negative				
	Positive	11987	513	0.96	0.55	0.70	
	Negative	9885	2615	0.21	0.84	0.33	
Unigram + Bigram		Correct Labels		Precision	Recall	F-Measure	83.36
		Positive	Negative				
	Positive	9409	3091	0.75	0.90	0.82	
	Negative	1069	11431	0.91	0.79	0.85	
Bigram + Trigram		Correct Labels		Precision	Recall	F-Measure	58.744
		Positive	Negative				
	Positive	12427	73	0.99	0.55	0.71	
	Negative	10241	2259	0.18	0.97	0.30	
Unigram + Bigram + Trigram		Correct Labels		Precision	Recall	F-Measure	83.336
		Positive	Negative				
	Positive	9423	3077	0.75	0.90	0.82	
	Negative	1089	11411	0.91	0.79	0.85	

Figure 4.5 shows a comparative analysis of accuracy obtained using different Naive Bayes based n-gram techniques.

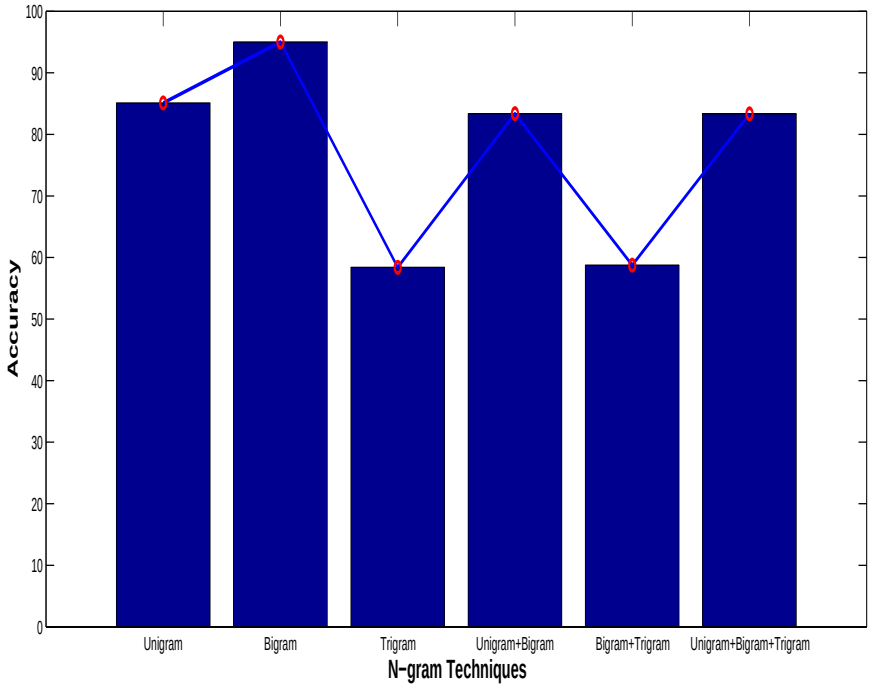

Figure 4.5: *Comparison of Accuracy values of different n-gram technique using SGD*

As illustrated in Table 4.4 and Figure 4.5, it can be analyzed that the accuracy obtained using unigram is better than the values obtained using bigram and trigram. In SDG method, the gradient is estimated on single randomly picked reviews using learning rate to minimize the risk. In unigram, a single word is randomly picked to analyze, but in bigram and trigram both the combination of the words adds noise, which reduces the value of accuracy. Thus, when the bigram and trigram model is combined with other model, their less accuracy value affects the accuracy of the total system.

4.6 Performance Evaluation

The comparative analysis based on results obtained using proposed approach to that of other literatures using IMDb dataset and n-gram approaches are shown in Table 4.5.

Pang et al., have used machine learning algorithm viz., Naive Bayes, Maximum Entropy method, and Support Vector Machine method using n-gram approach of unigram, bigram and combination of unigram and bigram. Salvetti *et al.* and Beineke *et al.* have implemented the Naive Bayes method for classification; but only the unigram approach is used for classification. Mullen and Collier, have proposed Support Vector machine method for classification; with unigram approach only. Matsumoto *et al.* have also implemented the Support Vector Machine for classification and used the unigram, bigram, and combination of both i.e., unigram and bigram for classification.

Table 4.5: Comparative result of values on ‘‘Accuracy’’ result obtained with different literature using IMDb Dataset and ngram approach

Method		Pang et al. [8]	Salvetti et al. [34]	Beineke et al. [35]	Mullen & Collier [36]	Matsumoto et al. [56]	Proposed Approach
Naive Bayes Classifier	Unigram	81.0	79.5	65.9	⊗	⊗	83.65
	Bigram	77.3	⊗	⊗	⊗	⊗	84.06
	Trigram	⊗	⊗	⊗	⊗	⊗	70.53
	Unigram + Bigram	80.6	⊗	⊗	⊗	⊗	86
	Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	83.82
	Unigram + Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	86.23
Maximum Entropy	Unigram	80.4	⊗	⊗	⊗	⊗	88.48
	Bigram	77.4	⊗	⊗	⊗	⊗	83.22
	Trigram	⊗	⊗	⊗	⊗	⊗	71.38
	Unigram + Bigram	80.8	⊗	⊗	⊗	⊗	88.42
	Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	82.94
	Unigram + Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	83.36
Support Vector Machine	Unigram	72.9	⊗	⊗	86.0	83.7	86.97
	Bigram	77.1	⊗	⊗	⊗	80.4	83.87
	Trigram	⊗	⊗	⊗	⊗	⊗	70.16
	Unigram + Bigram	82.7	⊗	⊗	⊗	84.6	88.88
	Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	83.63
	Unigram + Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	88.94
Stochastic Gradient Descent	Unigram	⊗	⊗	⊗	⊗	⊗	85.11
	Bigram	⊗	⊗	⊗	⊗	⊗	62.36
	Trigram	⊗	⊗	⊗	⊗	⊗	58.40
	Unigram + Bigram	⊗	⊗	⊗	⊗	⊗	83.36
	Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	58.74
	Unigram + Bigram + Trigram	⊗	⊗	⊗	⊗	⊗	83.36

⊗ indicate that the algorithm is not considered by the author in their respective paper

In this present chapter, four different algorithms viz., Naive Bayes, Maximum Entropy method, Support Vector Machine, and Stochastic Gradient Descent using n-gram approaches like unigram, bigram, trigram, unigram+bigram, bigram+trigram, and unigram+bigram+trigram are carried out. Result obtained using n-gram approach is observed to be better than the result available in the literature where both IMDb dataset and n-gram approach are used.

4.6.1 Managerial Insights Based on Result

The managerial insight based on the obtained result can be explained as follows:

- It is a practice that, sales managers obtain the feedback from customers or users on the product after it is used, in the form of reviews or blogs.
- The proposed approach classifies the reviews into either positive or negative polarity; hence the classes identified are able to guide the managers properly in order to sustain the market competition.

4.7 Summary

In this chapter, the n-gram approach for classification of the reviews are carried out on IMDb dataset. Four different machine learning techniques viz., NB, SVM, ME, and SGD are used for classification of reviews into two different classes. From the obtained result, it can be analyzed that the n-gram with lower value of 'n' i.e., unigram and bigram shows comparatively better result in case of all MLTs. But when the value of 'n' increases the accuracy result start decreasing i.e., in case of all MLTs, the accuracy result obtained using trigram is found to be less in comparison with those of unigram and bigram. the complexity of the proposed algorithms are as follows:

- Naïve Bayes: O(n * m)
- SVM: O(n * m)
- Maximum Entropy : $O(n * d^2)$
- Stochastic Gradient Descent (SGD) : $O(m * n^2 * logn)$

where, 'n' is the number of reviews present and 'm' is no of classes, in this case the value of 'm' is two as positive and negative classes are considered, 'd' in the scattered between the classes.

As sentiment analysis is concerned with analysis of text documents where each word, after removal of stop words and other unwanted information, is considered as a feature for analysis. But when the list of features turn out to be very large and confusing, they need to be considered carefully. Thus, in next chapters an attempt is made to select the best features and based on these features, the analysis is taken for consideration.

Chapter 5

Document level Sentiment Analysis using Genetic Algorithm and Neuro-Genetic Algorithm

In this chapter, the classification of the reviews are carried out based on two different algorithms i.e., Genetic Algorithm (GA) and Neuro Genetic Algorithm (NeuroGA) i.e., the hybrid form of of Artificial Neural Network and Genetic Algorithm. When GA is considered, the text reviews are represented in the form of chromosomes and then classified accordingly. While applying NeuroGA, the best features are selected from the set of large features by implementing GA and then, Neural Network classifies the reviews based on the selected features. Different performance evaluation parameters such as precision, recall, F-measure, and accuracy are considered to assess critically the performance of the classifiers.
The rest of the chapter is organized as follows:
Section 5.1 provides a brief introduction to the proposed approach and the contribution of the chapter. Section 5.2 Indicates the motivation for the proposed approach. Section 5.3 discusses about different methodologies used to transform text data into numerical vectors, classification techniques, and performance evaluation parameters. Section 5.4 highlights the proposed approach. Section 5.5 compares the performance of the proposed approach with present literatures. Finally, Section 5.6 summarizes the chapter.

5.1 Introduction

In document level sentiment analysis, each document is considered as a single unit for analysis. On the basis of this analysis, the document is classified into either a positive or negative group. The contribution of this chapter can be stated as follows:

i. The polarity dataset [33] is considered for sentiment classification using 10 fold cross validation technique, which contains 1000 positive and 1000 negative reviews. Thus, 900 positive and 900 negative reviews are considered for training and rest 100 positive and 100 negative reviews are considered for testing purpose.

ii. Each text review is represented as a vector of size equal to the dictionary size. The

dictionary size is equal to the total number of features or words present in the total text reviews. A Best Gene Vector (BGV) is generated with random occurrence of 1s and 0s of the size equal to dictionary size. The text reviews are represented in the form of chromosome by performing ‘‘logical AND’’ operation with the BGV.

iii. After representing all the reviews in the form of chromosome, the ‘‘selection’’ operation of GA is performed. The K-Nearest Neighbor (KNN) algorithm is considered to select the suitable or fit reviews. The reviews, which are classified correctly are considered as fit and the rest are treated as unfit, needing further processing. This process is carried out for ’n’ number of iterations to obtain the maximum number of fit chromosomes. In this chapter, maximum iterations (MAXITER) are assumed to be 1000.

iv. These fit chromosomes are then considered as an input to Artificial Neural Network (ANN) classifier, which classifies them to obtain the desired number of classes.

5.2 Motivation for the proposed approach

The Section 2.4 discusses about Sentiment Classification using hybrid MLTs and the Table 2.3 provides a comparative analysis of the approach used by various authors in literature. The comparative analysis provides clues to work on some possible research areas which can be extended further. The following aspects have been considered for carrying out research.

i. A number of authors have used part-of-speech (POS) tags or count of the occurrence of the word as a criteria for feature selection. But it is observed that the POS tag for a word is not fixed and it changes as per the context of their use. For example, the word ‘book’ can have the POS as ‘noun’, when used as reading material, where as in case of ‘‘ticket booking’’ the POS is a verb. Again the occurrence of a particular word mainly depends upon the author’s writing style. Thus, it may not be suitable for using feature selection. Hence, in this chapter, the sentiment values of the words are considered as features and the feature selection process is carried out at the document level.

ii. Different authors have used various machine learning techniques for both feature selection and classification purpose. Thus, it it may be observed that the shortcoming of the techniques bias the final classification result. To solve this issue, in this chapter, GA is considered for selection of features where as ANN is used for sentiment classification. Hence, any particular algorithm may not have any bias to affect the result of classification.

iii. Most of the machine learning algorithms work on the data represented in the form of matrix having numbers as its elements. But the sentiment data are always in text format. So, the textual data need to be converted to numerical values and they can be arranged in a matrix form. Different authors have considered TF or TF-IDF methods to convert

the text into a matrix having numbers as elements. But in this chapter, as binary GA is considered for classification, the countvectorizer (CV) method is used as data often it represent the occurrence of the feature with '1' and the non-occurrence as '0'.

iv. The sentiment analysis is carried out on text data where each word of the text review is considered as a feature. Thus, there are occurrences where the number of features may turn out to be very large and they might affect result of the analysis. In order to handle this situation, feature selection method needs to be carried out, in order to find the best features, acting as representative for whole dataset. This feature selection process not only saves the time for computation but also improves the accuracy. In this chapter, the GA is applied as feature selection for polarity dataset.

5.3 Methodology Adopted

This section discusses about the different methodologies adopted for classification of dataset based on movie reviews using hybrid machine learning techniques.

5.3.1 Types of sentiment classification

According to the Section 3.3.1, there exists two types of classification technique i.e., binary and multi-class sentiment analysis. It is observed that a good number of researchers have used binary classification technique for sentiment classification. Hence, it is adopted in this chapter for classification of movie reviews.

5.3.2 Transformation of Text Data into Numerical values / matrix

Since the sentiment reviews are in the text format, the MLTs need the data in the form of numerical vectors only for classification. The Section 3.3.2 discusses about the different transformation techniques for converting the text reviews into numerical vectors. In this chapter, the binary GA is used for classification which is based on the principle of either presence or absence of a particular feature. The CV technique is concerned with the presence or absence of the features and represents it with '1' and '0' respectively.

5.3.3 Dataset Used

For the purpose of case study, the polarity movie review dataset is considered for document level sentiment analysis. The polarity dataset consist of 1000 positive reviews and 1000 negative label reviews [33]. Though the database contains both negative and positive reviews, it is not partitioned for training and testing. Thus, in order to perform the classification, the 10 fold cross validation method is being used in this dataset. In this chapter, for training purpose 90% of datat i.e., 900 reviews are considered for training purpose and 10% i.e., 100 reviews are considered for testing.

5.3.4 Application of Machine Learning Techniques

In this chapter, the following two machine learning techniques are used.

1. **Genetic Algorithm (GA):** The basic principles of GA was first proposed by J.H. Holland [141]. GA simulates the natural population process with a population of ‘‘individuals’’, where each one represents a possible solution to a given problem . The application of GA can be explained in the form of flow-chart as in Figure 5.1 [142].

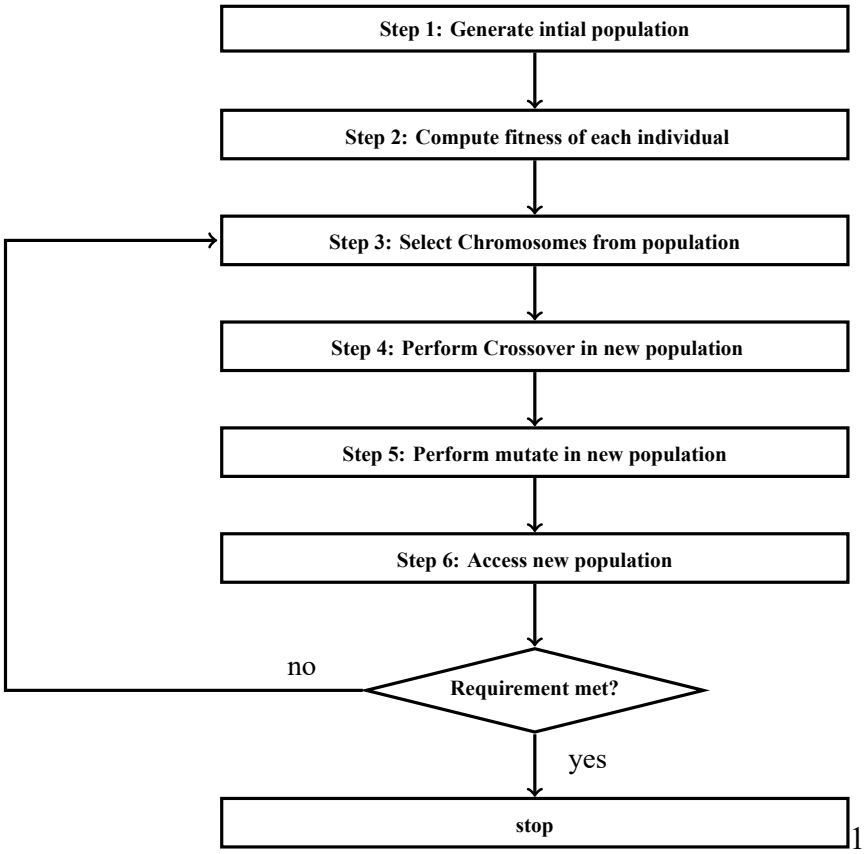

Figure 5.1: *Proposed approach for Classification using GA classifier*

(a) In case of GA analysis, the text data need to be represented as chromosome.

(b) A fitness function is considered for each element in the population. In this study, the k Nearest Neighbor (KNN) technique is used to find out the fitness value of each element. The elements being classified correctly are considered as fit chromosomes and the rest, which can not pass fitness test are considered to be unfit.

- K Nearest Neighbor: The KNN algorithm is used in this chapter as fitness function to find out the fit chromosomes among all. The process of KNN can be explained as follows [143]:
  - For a given document d, the KNN finds the nearest neighbors of d.

- The similarity score of each document to the test document is considered as weight of the class.
- Then the weighted sum of KNN is calculated as follows [144]:

$$Score(d, c_j) = \sum_{d_i \in KNN(d)} sim(d, d_i)\ y(d_i, c_j) \tag{5.1}$$

where
$Score(d, c_j)$ is the score of the candidate category c_j with respect to d;
$sim(d, d_i)$ is the similarity between d and the training document d_i;
$y(d_i, c_j) \in 0, 1$ is the binary category value of the training document d_i with respect to c_j
($Score(d, c_j) = 1$ indicates document d_i as part of category c_j, or $Score(d, c_j) = 0$).

(c) After identifying chromosomes which are fit, again a pair of chromosomes are selected from the group of fit elements and different GA operators are applied on them.

(d) Crossover process takes two different chromosomes and cuts their chromosome string at any random position to produce two head and two tail segments. The tail segments are then swapped over to produce full length chromosome. Cross over is not applied to each pair of chromosomes; rather few pairs are chosen at random to perform this operation.

(e) Mutation is carried out on each child, produced after the crossover. This operation randomly changes the genes of the chromosome.

(f) The process of selection, crossover and mutation process is carried out until the desired result is obtained.

2. **Artificial Neural Network (ANN):** Neural network used for classification can be represented as a mapping function such as

$$F : A^d -> A^m \tag{5.2}$$

where
'd' the dimensional input is provided to network; and
'm' vector output is obtained with classification result.
The Figure 5.2 shows the structure of a neural network. The input layer of neural network consists of 'd' neurons representing 'd' pieces of input signal (Independent variable). The number of neurons in the hidden layer are chosen by the user. The output layer consists of 'm' number of neurons (considered as dependent variables) [145].

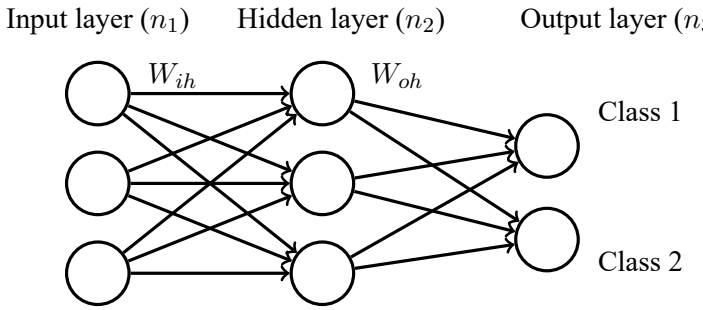

Figure 5.2: A typical neural network

In the input layer, the state of each neuron is determined by input variable. For other neurons the state of neurons are evaluated using values associated with previous neurons as per following equation:

$$a_j = \sum_{i=1}^{I} X_i W_{ji} \tag{5.3}$$

where

a_j is the net input of neuron j;

X_i is the output value of $neuron_i$ in previous layer;

W_{ji} is the weight factor of the connection between neuron i and neuron j.

The neuron's activity is very often determined with the help of a sigmoid function as mentioned below:

$$g(a_j) = \frac{1}{1 + exp^{-a_j}} \tag{5.4}$$

In back propagation technique, each iteration tries to minimize the error. The adjustment of weight is initiated from output layer to input layer [146]. Error correction is carried out using following function:

$$\Delta W_{ji} = \eta \delta_i F(a_i) \tag{5.5}$$

where

ΔW_{ji} is the adjustment of weight between neuron j and i;

η is the learning rate;

δ_i depends on the layer;

$F(a_i)$ is the output of network 'i'.

The training process is carried out till the error is minimized. After the completion of training process, the performance of NN is tested using the input data and the performance is measured using the elements in confusion matrix.

5.3.5 Parameters used for Performance Evaluation

As discussed in Section 3.3.6, confusion matrix is used for evaluation of the performance of the MLTs. The Table 3.1 shows the confusion matrix, which includes elements like true positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Using this elements, few parameters like precision, NPV, recall, TNR, F-measure and accuracy are found out to evaluate the performance of the MLTs.

5.4 Proposed Approach

In this chapter, the polarity dataset [33] has been considered for analysis. As the reviews are not separated into test and training reviews, 10 fold cross validation technique is used to evaluate the performance of the classifier. Thus, 900 positively labeled as well as negatively labeled reviews are considered for training and rest 100 positively labeled reviews and 100 negatively labeled reviews are considered for testing purpose. The detailed procedure is described in following two sections namely Classification using Genetic Algorithm and Classification using Neuro-GA discussing about the classification approach used in case of GA and NeuroGA respectively.

5.4.1 Classification using GA

The figure 5.3 explains the step-wise procedure for the classification of movie reviews using Genetic Algorithm. The explanation of steps is as follows.

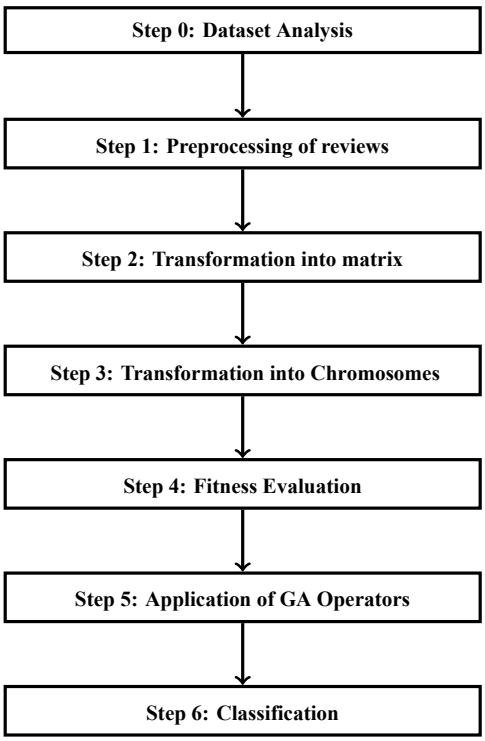

Figure 5.3: *Proposed approach for Classification using GA classifier*

1. The reviews need to be pre-processed for removal of unwanted noise from the data and for normalization. The steps are as follows:

   - **Removal of Stop Words**: Stop words do not contribute towards the sentiment; thus they may be removed.
   - **Removal of Numeric and Special Character**: In the text reviews, there may be some numeric (1, 254, ..., 50 etc.) and special characters (@, #, $, % etc.) present, which do not contribute towards the sentiment, but may create confusion while transforming them into numeric vector form. Hence, they may be removed.
   - **Removal of URL and HTML tags**: As the reviews are fetched from different sites, so they may contain URLs and HTML tags. URLs and HTML tags do not contribute towards the sentiment; therefore may be removed.
   - **Lowering of Case**: The reviews do not contain the all words in the same format. Thus, it is needed to convert them in to a uniform case which help the analysis process. In this chapter, all the words converted into lower case to a maintain an uniformity among all words which help to remove redundancy between the words.
   - **Stemming**: The stemming process is carried out i.e., the process of extracting the root word from the given word (e.g., the root word for reading and readable will be read).

2. After preprocessing of text reviews, they are transformed into numeric vectors of size *n* (Size of Dictionary). The CV technique is used to do the same. It converts the text reviews into a vector of 0's and 1's.

3. A Best Gene Vector (BGV) of size *n* (Size of Dictionary) consisting of 0's and 1's is randomly chosen. Presence of 1's in the BGV indicate that the genes/features are significant ones and 0's indicate otherwise. By controlling number of occurrences of 1's, number of significant genes/features to be filtered out can be specified. Using the BGV, each review is then transformed into chromosome to be used in GA. To obtain i-th chromosome, logical AND operation is carried out between vector representation of the i-th review with BGV.

4. Using the KNN (K-Nearest Neighbor), each chromosome is classified into a positive and negative group. If the classification is of true-positive or true-negative category, then the chromosome is treated as fit and if classification is of false positive or false negative then the Chromosome is unfit.

5. Chromosomes with highest accuracy / fitness values are selected. After selection, crossover is performed with a probability of 0.6. This reduced probability helps in

transformation of fit genes without any alteration. To perform crossover, a random crossover point is chosen. Then heads and tails are swapped to generate two new chromosomes for new population. Then mutation is also performed with probability of 0.001 at a random place in the chromosome and then flipped. This process is repeated on the new population. The process is repeated until the BGV converges or MAXITER has reached. In this chapter, the value of MAXITER is considered as 1000.

6. The 1's in BGV indicates the best genes or the best features that should be used for classification. This reduces the size of vocabulary from 25956 to 3394.

The following table 5.1 shows the result obtained using GA classification

Table 5.1: Result obtained using GA classifier

	Precision	Recall	Accuracy
Positive	0.928	0.933	0.93
Negative	0.933	0.928	

5.4.2 Classification using NeuroGA

The Figure 5.4 explains the step-wise procedure for the classification of movie reviews using NeuroGA.

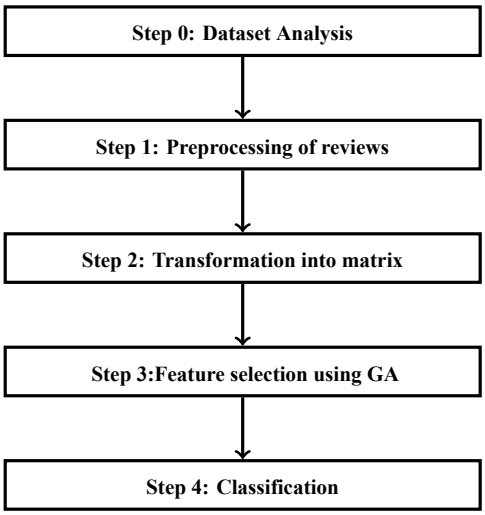

Figure 5.4: *Diagrammatic view of the proposed approach using Neuro - GA classifier*

In **Classification process using Neuro - GA**, the best features which are found out using Genetic Algorithm to perform classification, have been reconsidered.

- Steps 0 to 2 of earlier section, i.e., Section 5.4.1 are repeated as in **Classification using Genetic Algorithm**.

- Also its Step 3 is the step for selection of best features which is outcome of previous method.
- After the feature selection activity is carried out, the input data is considered for testing. For classification, 10 fold cross validation is being used i.e., 90% of the reviews are used for training i.e., already being done. Then for the testing of the rest, 10 % of reviews are carried out. The result is being analyzed using various performance evaluation parameters like precision, recall and accuracy. The number of hidden neurons depend upon the choice of user. So, in this chapter, the number of hidden nodes are considered to be in order of 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000 and 5000. The input matrix considered for analysis is of size **1800 x 3394**, and the output has two neurons i.e., either positive or negative. Table 5.2 shows the result obtained using NeuroGA classification

Table 5.2: Result obtained using different number of hidden nodes in ANN

Number of hidden neurons	Precision		Recall		Accuracy
	Positive	Negative	Positive	Negative	
100	.74	.86	.89	.77	.789
200	.83	.84	.84	.83	.83
300	.869	.87	.863	.87	.87
400	.89	.87	.88	.89	.88
500	.83	.84	.84	.82	.83
600	.9	.91	.91	.9	.911
700	.91	.92	.92	.91	.91
800	.928	.933	.933	.928	.93
900	.94	.938	.95	.94	.94
1000	.958	.967	.967	.958	**.963**
2000	.931	.947	.948	.93	.939
3000	.934	.931	.931	.934	.933
4000	.92	.925	.926	.93	.92
5000	.909	.914	.914	.909	.911

From the Table 5.2 and Figure 5.5, it is observed that the accuracy of the proposed system is in an increasing mode, up to a specific number of hidden nodes i.e., 1000 hidden nodes in this case. But after this stage, the accuracy of the system decreases due to over fitting of the machine learning algorithm; thus the hidden nodes are kept up to 5000 in this experiment.

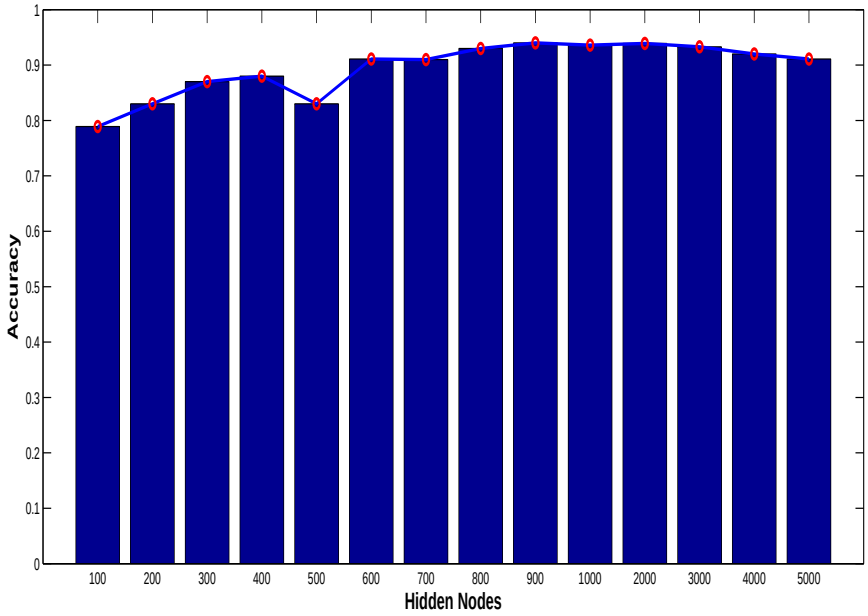

Figure 5.5: *Comparison of accuracy values of using different hidden nodes for ANN*

5.5 Performance Evaluation

Comparative analysis on accuracy values using the proposed approach with other approaches as available in literature where same polarity dataset, has been considered and presented in Table 5.3.

Table 5.3: Comparative analysis of results with different literature using polarity dataset

Authors	Approach used	Accuracy
Pang and Lee [33]	Naive Bayes (NB), SVM	NB = 0.864 , SVM = 0.872
Whitelaw et al. [138]	SVM	SVM = 0.902
Matsumoto et al. [56]	SVM	SVM = 0.937
Aue and Gamon [135]	SVM	SVM = 0.905
Read [136]	NB, SVM	NB = 0.789 , SVM = 0.815
Kennedy and Inkpen [137]	SVM	SVM = 0.862
Mores et al. [42]	NB, SVM, ANN	NB = 0.803, SVM = 0.841, ANN = 0.865
Proposed approach	GA and NeuroGA	**GA = 0.93 , NeuroGA = 0.963**

From the Table 5.3 and Figure 5.6, it is observed that the proposed approach i.e., the NeuroGA approach yields better result as compared to those of similar approaches, available literatures. The feature selection process selects the best features and then these selected

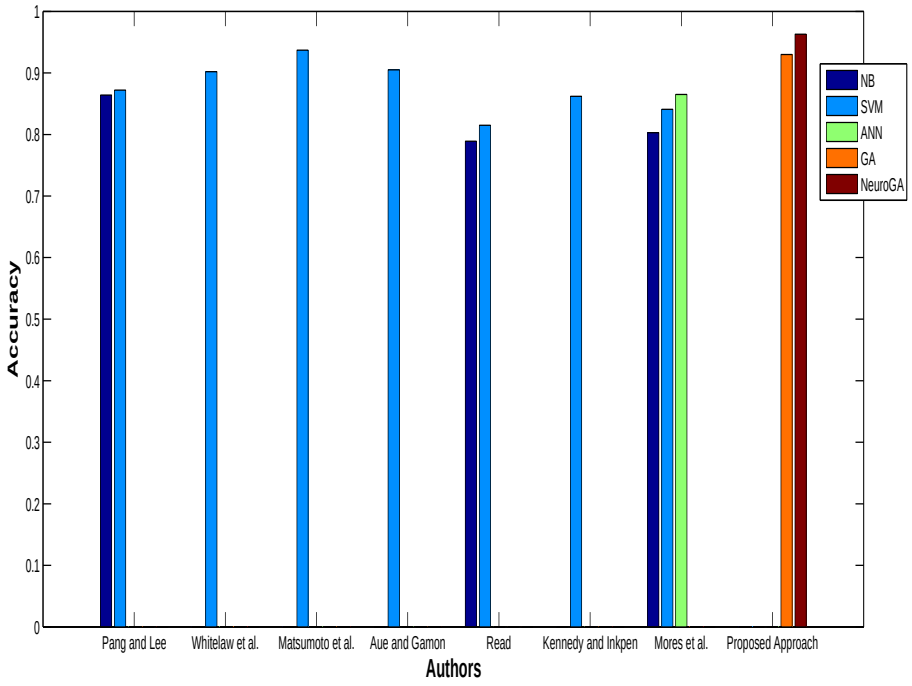

Figure 5.6: *Comparison of accuracy values of obtained by different authors on Polarity dataset*

features are considered as input to ANN. This process helps improve the accuracy.

5.5.1 Managerial Insights Based on Result

The following points indicate various managerial insights based on the obtained result.

- The proposed approach classifies the reviews into either positive or negative polarity; hence is able to guide the managers properly by informing them about the shortcoming or good features of the product which the decision makers definitely consider to implement the sentiments for quality improvement purpose in order to sustain the market competition.

5.6 Summary

In this chapter, the hybrid machine learning technique is proposed using GA and ANN. The output i.e., the fit chromosomes obtained after analysis of polarity dataset are then given input to NeuroGA for analysis. The hidden nodes of the ANN are kept on changing until to best accuracy value is obtained. In this chapter, the best accuracy is obtained for hidden value of 1000 neurons. Apart from that, the GA also classifies the reviews to obtain a classification accuracy of 93%.

In this chapter, the best chromosomes are selected and based on that analysis is done. But still all the words are considered. Thus, in the next chapter, an attempt is made to obtain

the sentiment values for each word and then the words with higher sentiment values which are either positive or negative by nature, are considered for further analysis using MLTs.

Chapter 6

Document level Sentiment Classification using Feature Selection Technique

During the process of sentiment analysis, each word is considered as a feature, and based on these features analysis is carried out. But in most of the cases the collection of these words are very big; thus, in this chapter the Support Vector Machine (SVM) method is implemented as feature selection method based on the sentiment values of the words. These selected words are then given input to ANN for classification. By changing the hidden nodes of ANN, the accuracy of the proposed system is found out. The results are compared with the result available in the existing literatures. The rest of the chapter is organized as follows:
Section 6.1 provides a brief introduction to the proposed approach and the contribution of the chapter. Section 6.2 Indicates the motivation for the proposed approach. Section 6.3 discusses about different methodologies used to transform text data into numerical vectors, classification techniques, and performance evaluation parameters. Section 6.4 highlights the proposed approach. Section 6.5 compares the performance of the proposed approach with the result available in present literature. Finally, Section 6.6 summarizes the Chapter.

6.1 Introduction

In this chapter, two different movie review datasets i.e., IMDb [32] and Polarity [33] are considered for analysis. SVM technique is used to find out the sentiment value of each word and the best features are then given as input to ANN for classification. The contribution of this chapter can be stated as follows:

- The IMDb dataset has separate data for training and testing but as polarity dataset has no such separation. Hence, 10 fold cross validation technique is used for classification.
- After removal of stop words and unwanted information from the training data in both datasets; the rest of the words are considered as features for sentiment classification. The sentiment values of each word are calculated using SVM; then, for feature selection purpose, the words having sentiment values, higher than mod(0.009) are considered.

- These selected features / words are then given input to ANN, which tests the testing data based on selected features and classifies the review dataset into either positive or negative polarity.
- During the analysis using ANN, the number of hidden nodes is kept on changing to find out the best possible solution. With the help of confusion matrix and other performance evaluation parameters like ‘‘precision’’, ‘‘recall’’, ‘‘F-measure’’, and ‘‘accuracy’’, the performance of the proposed approach is compared with the results available in literatures.

6.1.1 Motivation for the Proposed Approach

The Section 2.5 discusses about Sentiment Classification using feature selection approach and the Table 2.4 provides a comparative analysis of those papers. These information help to identify some possible research areas which can be extended further. The following aspects have been considered for carrying out further research.

i. It is observed that a good number of authors have used part-of-speech (POS) tags or count of the occurrence of the word as a criteria for feature selection. But it is observed that the POS tag for a word is not fixed and it changes as per the context of their use. Again the occurrence of a particular word mainly depends upon the author’s writing style. Thus, it may not be suitable for using feature selection. Hence, in this chapter, the sentiment values of each word is calculated and feature selection is carried out on the basis of this sentiment values.

ii. Different authors have used same machine learning technique for both feature selection and classification. Thus, it is found out that the shortcoming of the same technique may bias the final classification result. To solve this issue, in this chapter, SVM is used as feature selection where as ANN is used for sentiment classification. Hence, the bias of any algorithm does not affect the result of classification.

iii. Most of the machine learning algorithms work on the data represented as matrix of numbers. But the sentiment data are always in text format. So, it needs to be converted to numerical form and can be arranged in matrix. Different authors have considered TF or TF-IDF to convert the text into matrix of numbers. But in this chapter, in order to convert the text data into matrix of numbers, the combination of TF-IDF and CountVectorizer technique has been applied.

6.2 Methodology Adopted

This section discusses about the different methodologies adopted for classification of sentiment reviews.

6.2.1 Types of sentiment classification

According to the Section 3.3.1, there exists two types of classification techniques i.e., binary and multi-class sentiment analysis. It is observed that quite a good number of authors have used binary classification technique which is adopted in this chapter for classification of movie reviews. In case of multi-class classification, the reviews are classified into five different classes according to Section 3.3.1. But, there is no specific boundary specified for the reviews to belong to either strong positive or positive. Thus, it solely depends upon the review to specify the levels. it may be noted that for supervised learning, the training data need to be represented in multi-class. So, by going through that knowledge the MLTs train them and finally predict the polarity of the reviews. Thus, the binary classification is mainly used by different authors for classification.

6.2.2 Transformation of Text Data into Numerical values / matrix

The sentiment reviews are in the text format, where as the MLTs need the data in the form of numerical vectors for classification. Thus, the reviews need to be transformed into numerical form. The Section 3.3.2 discusses about the different transformation techniques for converting the text reviews into numerical vectors. The two transformation mechanisms i.e., CV and TF-IDF are used simultaneously in this chapter for representation of the text as numerical vector.

6.2.3 Dataset used

In this chapter, for the sentiment classification of movie reviews, both aclIMDb [32] and Polarity dataset [33] are used. This IMDb dataset contain separate data for training and testing. Both training and testing dataset contain 12500 positive and 12500 negative reviews respectively. While the polarity dataset contains 1000 positively and 1000 negatively labeled reviews. So, for classification 10 fold cross validation technique is used i.e., 90% of the reviews which is 900 positively and 900 negatively labeled reviews are considered for training and rest 100 positively and 100 negatively labeled reviews are considered for testing.

6.2.4 Machine Learning Technique Used

After preprocessing of the text reviews, they need to be transformed into numerical vectors which are then given as input to different machine learning techniques for classification. In this chapter, two different machine learning techniques i.e., SVM and ANN used for classification. The detailed implementation of the SVM technique is already discussed in Section 3.3.5. SVM technique is used for classification in Section 3.3.5; but in this chapter, SVM technique is used to find out the sentiment score for each word after removal of unwanted information. These sentiment scores form the selection criteria for feature

selection. The detailed information about the implementation of ANN is discussed in Section 5.3. In Chapter 5, the GA selects the fit chromosomes and then these fit chromosomes are considered as input to ANN, on the basis of which ANN classifies the reviews with changing hidden nodes. In this chapter, the SVM technique selects the best features on the basis of their sentiment score and then selects the features with higher sentiment values. These higher sentiment score words are then given as input to ANN which then classify the reviews by varying the number of hidden nodes to find out the right kind of classification.

6.2.5 Parameters used for Performance Evaluation

As discussed in Section 3.3.6, confusion matrix is used for evaluation of the performance of the MLTs. The Table 3.1 shows the confusion matrix, which includes terms like true positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Using this terms, some parameters like precision, NPV, recall, TNR, F-measure and accuracy are calculated to evaluate the performance of the MLTs.

6.3 Proposed Approach

In this study, IMDb and polarity datasets are considered for analysis. The datasets are preprocessed in order to remove the stop words from dataset. The processed textual data are then used to obtain a matrix of numerical vectors using vectorization techniques. Further, the dataset go through a feature selection step, which selects the features depending upon some conditions. The dataset is further classified based on the feature selected. Stepwise elaboration of the approach is described in figure 6.1.

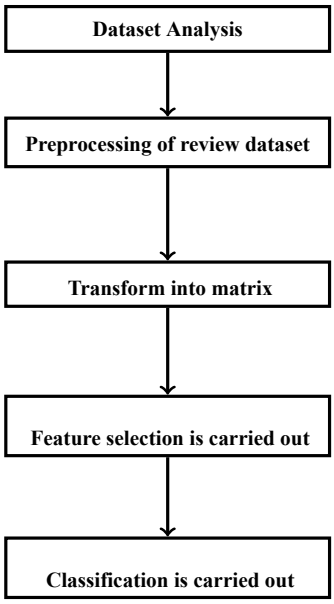

Figure 6.1: *Diagrammatic view of the proposed approach*

Step 1. The two datasets IMDb[32] and polarity [33] are used for classification.

Step 2. The text reviews consist of absurd information which needs to be removed from the original reviews before it is considered for classification. The details of these absurd information are discussed in Section 3.4.

Step 3. After the preprocessing of text reviews, they need to be arranged into a matrix form of numeric vectors. The algorithms for conversion of text file to numeric vectors are as follows:

- CV: It converts the text reviews into a matrix of token counts. It implements both tokenization and occurrence counting.
- TF-IDF: It suggests the importance of the word to the document and to the whole corpus. Term frequency informs about the frequency of a word in a document and IDF informs about the frequency of the particular word in whole corpus.

Step 4. After the text reviews are converted to numeric vectors, these information are then considered for the process of feature selection using SVM algorithm. The steps for feature selection for different datasets are shown as follows:

(a) The IMDb dataset has 12500 positive reviews and 12500 negative reviews for training and the same amount of reviews present for testing purpose. For the purpose of feature selection, only the training data is considered.

The polarity dataset has 1000 positive and 1000 negative reviews. As 10 fold cross validation technique is used for classification, 900 positive and negative reviews are considered as input for feature selection. As the step of feature selection is mainly carried out on training data.

(b) After the preprocessing stage, when the unwanted words or information are removed, rest of the words are then considered as feature. After preprocessing, the total number of features obtained from IMDb dataset is 159438 and from polarity dataset is 25579 features .

(c) Then a matrix is generated, where the row specifies the file and the column specifies the feature with its occurrence.

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & ..... & x_{1n} \\ x_{21} & x_{22} & x_{23} & ..... & x_{2n} \\ ... & ... & ... & ..... & ... \\ x_{m1} & x_{m2} & x_{m3} & ..... & x_{mn} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ ... \\ \alpha_n \end{bmatrix}$$

Each element x(i,j) represents the occurrence of feature j in review i and α_i is a random variable multiplied with the feature.

$$\begin{bmatrix} \alpha_1 x_{1,1} + \alpha_2 x_{1,2} + \alpha_3 x_{1,3} + \ldots\ldots\ldots + \alpha_{25579} x_{1,25579} \\ \alpha_1 x_{2,1} + \alpha_2 x_{2,2} + \alpha_3 x_{2,3} + \ldots\ldots\ldots + \alpha_{25579} x_{2,25579} \\ \alpha_1 x_{3,1} + \alpha_2 x_{3,2} + \alpha_3 x_{3,3} + \ldots\ldots\ldots + \alpha_{25579} x_{3,25579} \\ \ldots\ldots\ldots \\ \ldots\ldots\ldots \\ \ldots\ldots\ldots \\ \ldots\ldots\ldots \\ \ldots\ldots\ldots \\ \alpha_1 x_{2000,1} + \alpha_2 x_{2000,2} + \ldots\ldots\ldots + \alpha_{25579} x_{2000,25579} \end{bmatrix}$$

As the review is supervised, it may be identified as to whether the sum of the product of α_1 and $x_{i,j}$ is positive or negative. If the sum total of the product is positive then its review is considered to be of positive polarity or else it is of negative polarity. As the value of α_1 is considered at random sometime the polarities do not match, in that case another set of α_1 is considered. These α_1 values finally show the polarity value of the respective features.

(d) Thus for each feature, the polarity value is obtained; but all words do not affect the polarity of review in same order. In this present chapter, the features are selected whose sentiment value is greater than mod(0.009). Thus the set of features is reduced to 19729 from 159438 for IMDb dataset and to 3199 from 25579 for Polarity dataset.

Step 5. After the feature selection is complete, then the input data is considered for testing. For classification, the 10 fold cross validation process has been adopted i.e., 90% of the review are used for training i.e., already being done. Then for the testing of the rest, 10 % of reviews are carried out. The result is being analyzed using various performance evaluation parameters like precision, recall and accuracy. The number of hidden neurons depends upon the user to obtain the optimal value; so, in this chapter, the numbers of hidden nodes are considered to be order of 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000 and 5000. The input matrix considered for IMDb dataset is of size 25000 x 19729 and the output has two neurons i.e., either positive or negative. The input matrix considered for Polarity dataset is of size 1800 x 3199 and the output has two neurons i.e., either positive or negative.

The following Table 6.1 and Table 6.2 show the result obtained using ANN classification on IMDb dataset and Polarity dataset respectively.

Table 6.1: Result obtained using different number of hidden nodes on IMDb dataset

Number of hidden neurons	Precision		Recall		Accuracy
	Positive	Negative	Positive	Negative	
100	.67	.76	.73	.70	.714
200	.82	.83	.83	.83	.83
300	.84	.83	.83	.84	.84
400	.88	.86	.86	.88	.87
500	.88	.89	.89	.88	.884
600	.99	.91	.92	.99	**.95**
700	.92	.9	.9	.92	.91
800	.88	.88	.86	.88	.87
900	.79	.91	.9	.81	.851
1000	.89	.79	.81	.88	.84
2000	.88	.78	.89	.87	.83
3000	.82	.83	.83	.83	.83
4000	.67	.76	.73	.7	.713
5000	.81	.6	.67	.76	.705

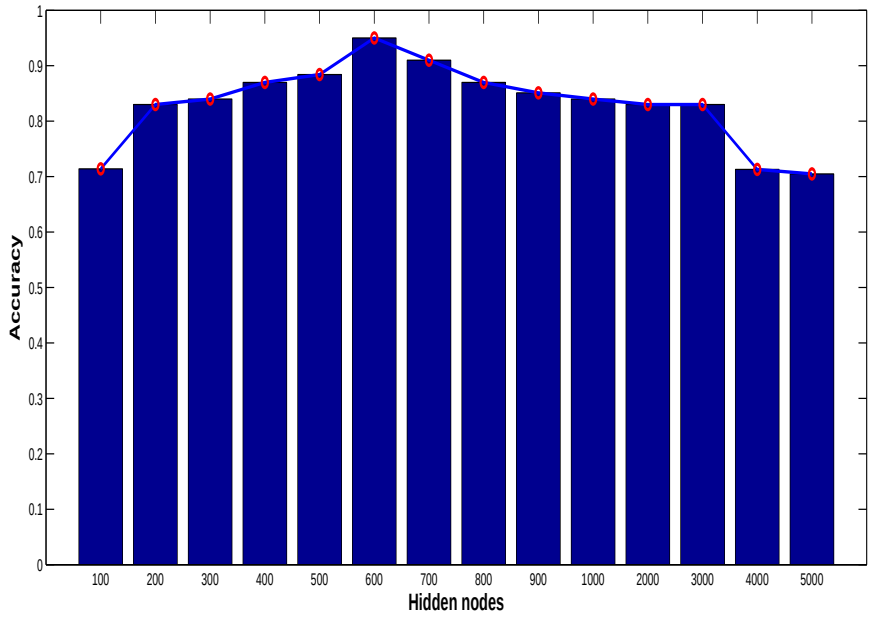

Figure 6.2: *Comparison of accuracy values of using different hidden nodes for ANN for IMDb dataset*

Table 6.2: Result obtained using different number of hidden nodes on Polarity dataset

Number of hidden neurons	Precision		Recall		Accuracy
	Positive	Negative	Positive	Negative	
100	.689	.714	.732	.67	.701
200	.869	.865	.863	.87	.867
300	.867	.857	.855	.869	.862
400	.902	.914	.915	.901	.908
500	.97	.958	.957	.97	**.964**
600	.958	.967	.967	.958	.963
700	.931	.947	.948	.93	.939
800	.934	.931	.931	.934	.933
900	.928	.933	.933	.928	.930
1000	.909	.914	.914	.909	.911
2000	.905	.906	.906	.905	.905
3000	.902	.899	.899	.902	.900
4000	.882	.866	.863	.885	.874
5000	.83	.84	.842	.828	.835

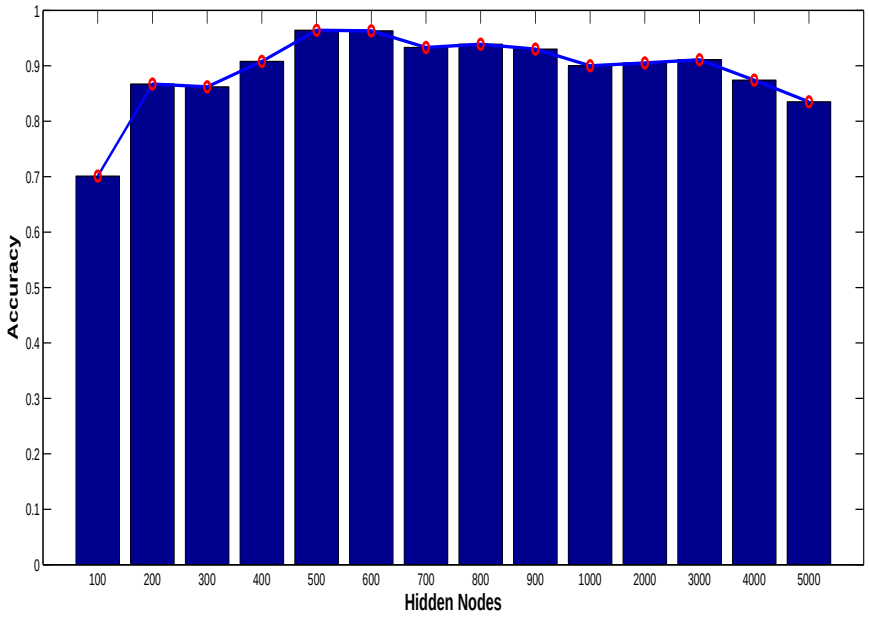

Figure 6.3: *Comparison of accuracy values of using different hidden nodes for ANN for polarity dataset*

From the Table 6.1 and Figure 6.2 it may be observed that the accuracy of the proposed system is in an increasing mode up to a specific number of hidden nodes i.e., for IMDb, it is 600 hidden nodes. But after these nodes, the accuracy of the system decreases due to over fitting of the machine learning technique. Similarly from Table 6.2 and Figure 6.3, it can be observed that for polarity it is 500 hidden nodes and after that the accuracy of system decreases due to over fitting of the machine learning technique.

6.4 Performance Evaluation

The Table 6.3 shows the comparison of accuracy values using the proposed approach with other approaches as available in literature using IMDb dataset and Table 6.4 shows the comparison of accuracy values using the proposed approach with other approaches as available in literature using Polarity dataset.

Table 6.3: Comparative result obtained using IMDb dataset

Author	Machine Learning Techniques Used	Accuracy
Pang et al. [8]	Naive Bayes (NB), SVM	NB = 0.815, SVM = 0.659
Salvetti et al. [34]	NB	NB = 0.796
Mullen and Collier [36]	SVM	SVM = 0.86
Beineke et al. [35]	NB	NB = 0.659
Matsumoto et al. [56]	SVM	SVM = 0.883
Proposed Approach	Hybrid of SVM and ANN	SVM + ANN = 0.95

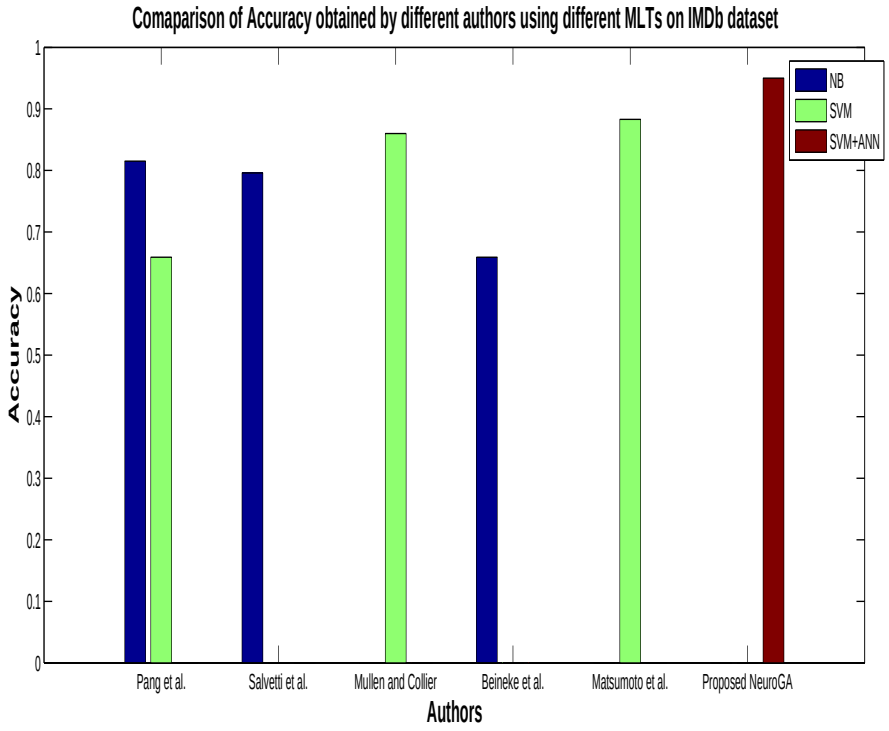

Figure 6.4: *Comparison of accuracy values of obtained by different authors on IMDb dataset*

Table 6.4: Comparative result obtained using Polarity dataset

Authors	Machine learning Techniques used	Accuracy
Pang and Lee [33]	Naive Bayes (NB), SVM	NB = 0.864 , SVM = 0.872
Whitelaw *et al.* [138]	SVM	SVM = 0.902
Matsumoto et al. [56]	SVM	SVM = 0.937
Aue and Gamon [135]	SVM	SVM = 0.905
Read [136]	NB, SVM	NB = 0.789 , SVM = 0.815
Kennedy and Inkpen [137]	SVM	SVM = 0.862
Moraes et al. [42]	NB, SVM,	NB = 0.803, SVM = 0.841, ANN = 0.865
Proposed approach	Hybrid of SVM and ANN	**SVM + ANN = 0.964**

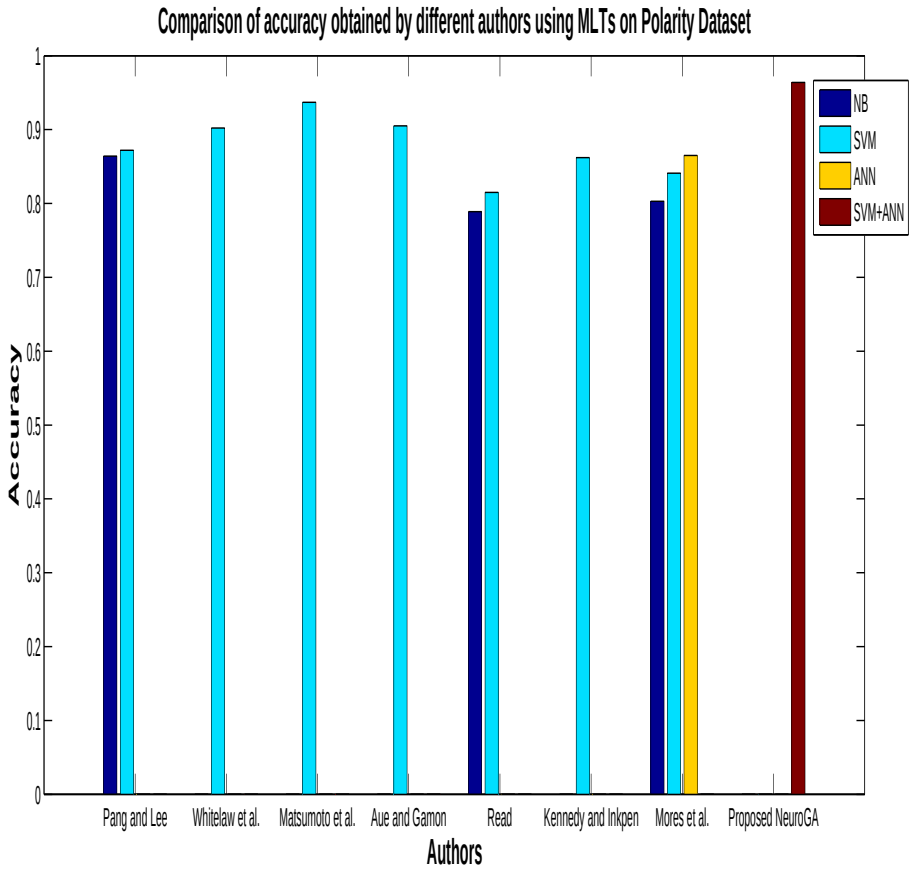

Figure 6.5: *Comparison of accuracy values of obtained by different authors on Polarity dataset*

From the Table 6.3 and Figure 6.4, and 6.3 and Figure 6.5, it may be observed that the proposed approach i.e., the hybrid approach of SVM and ANN shows much better result as compared to those of the other approaches. It may also be noted from the table that SVM yield comparative better results.

6.4.1 Managerial Insights Based on Result

The managerial insight based on the obtained result can be explained as follows:

- The reviews can be collected and given input to the proposed approach for qualitative decisions to be made for future.
- The proposed approach classifies the reviews into either of positive or negative polarity; hence is able to guide the managers properly for future decision making, to sustain the market competition.

6.5 Summary

In this chapter, the feature selection technique is adopted using SVM. The SVM technique finds out the sentiment values of each feature and based on the these values, it selects the features. These features are collected only from the training data. These are given as input to ANN. By varying the number of hidden nodes, the ANN finds out the classification of the reviews with best accuracy value. From Table 6.3 and Table 6.4, it is evident that the proposed method has shown a better classification result as compared to those of existing results.

Chapter 7

Sentiment Clustering using Unsupervised Machine Learning Technique

In this chapter, the sentiment analysis of the twitter tweets are carried out using unsupervised machine learning techniques. Four different machine learning techniques are used in the process of clustering and different performance evaluation parameters are used to evaluate the performance of the techniques.
The rest of the chapter is organized as follows:
Section 7.1 provides a brief introduction to the proposed approach and the contribution of the chapter. Section 7.2 Indicates the motivation for the proposed approach. Section 7.3 discusses about different techniques to transform text data into numerical vectors, clustering techniques, and performance evaluation parameters. Section 7.4 highlights the proposed approach along with the output obtained after implementation. Section 4.6 compares the performance of the proposed approach with present literatures. Finally, Section 4.7 summarizes the Chapter.

7.1 Introduction

Sentiment analysis is concerned with analysis of text reviews and generate a meaningful information from these reviews. In supervised machine learning techniques, the reviews are labeled into different classes; but collection of these reviews is a difficult task. Thus, in this chapter the unsupervised approach of the analysis is carried out for the reviews where the are not labeled and based on these reviews the analysis is carried out. The contribution of this chapter can be stated as follows:

- In this chapter, the Twitter dataset is considered for analysis. 42000 tweets related to politics are collected using Twitter API for the clustering of the reviews.
- For clustering of reviews, four different unsupervised machine learning techniques are applied namely K means, mini batch K-means, Affinity Propagation, and DBSCAN.
- Different performance evaluation parameters such as Homogeneity, Completeness, V-measure, Adjusted Rand Index, and Silhouette Coefficient are used to evaluate the

performance of the unsupervised machine learning techniques.

7.2 Motivation for the proposed approach

The Section 2.6 discusses about the unsupervised sentiment classification technique and Table 2.5 provides a comparative analysis of those papers. These paper helps to identify some possible research areas which can be extended further. The following aspects have been considered for carrying out further research.

- It is observed that a good number of authors in literature have considered the Twitter dataset for unsupervised sentiment analysis. People share their views about any topic using tweets. In this chapter, the Twitter tweets are collected using #indianpolitics. Twitter API is used to collected the tweets from the reviews. 42000 tweets are collected for the analysis.
- In supervised machine learning the reviews are labeled. Thus, sentiment classification is carried out to classify the reviews into different classes. But, in unsupervised sentiment classification technique the reviews are not labeled and thus, clustering approach is considered for analysis of the reviews. During the process of clustering, the reviews having the same properties are grouped into one cluster. In this chapter, the tweets are clusters in to two different clusters i.e. positive and negative.
- Like supervised machine learning technique, unsupervised MLTs also work on the reviews represented as matrix of numbers. But, the Twitter tweets are mainly in the text format. So, they need to be transformed into matrix of number for the analysis. Different authors have used either CV or TF or TF-IDF to transform the reviews into matrix of numbers. In this chapter, the CV and TF-IDF approach is combined together to obtain a numerical representation of the tweets.

7.3 Methodology Adopted

This section discusses about the different methodologies adopted for clustering of the Twitter tweets.

7.3.1 Sentiment Clustering

Clustering is the technique of dividing the data into meaningful and useful clusters [147]. The clustering process plans to discover natural grouping between data and thus, represent the analysis of the classes as a collection of document [148]. The clustering of the reviews is carried out on the basis of internal properties of reviews. Thus, there is no need of extracting any supervised information during the clustering process.

Clustering can be of two types i.e., hierarchical and partition clustering [149]. Hierarchical clustering deals with hierarchical decomposition of dataset i.e., if the clusters have sub-clusters then this clustering approach is used. But, partition cluster divides the dataset into non overlapping subsets. In binary sentiment analysis, there is no nested sub-cluster. Thus, partition clustering is more appropriate for sentiment clustering technique.

7.3.2 Transformation of Text Data into Numerical vector

The sentiment reviews are mainly in the text format and the MLTs need the data in the form of numerical vectors only for classification. The Section 3.3.2 discusses about the different transformation techniques for converting the text reviews into numerical vectors. The two approaches for transformation i.e., countvectorizer and TF-IDF are used simultaneously in this chapter for better representation of the text as numerical vector.

7.3.3 Dataset Used

People prefer to share their knowledge, views about any product through comments and views. Different social networking like facebook, Twitter sites help them to share their views. In this chapter, the tweets are collected from Twitter for analysis . The tweets can be collected from Twitter using following steps:

- The Twitter tweet are collected using Twitter API and R library called ‘‘twitterR’’.
- The twitterR package consist of two packages i.e., setup_twitter_oauth() and searchTwitter().
- The setup_twitter_oauth() package authenticate the user by using series of authentication code. Then, the tweets are collected from Twitter by using searchTwitter(). The Tweets are collected from Twitter based on number of tweet, tweet category and duration.
- In this chapter, in order to analysis the tweets, #indianpolitics is considered. 1500 tweets are collected per day for 28 days, thus, in total 42000 tweets are collected for analysis.

7.3.4 Machine Learning Technique Used

After the transformation of text reviews to numerical vectors, those are as given as input to the MLTs for classification purpose. In this chapter, four different MLTs are used for clustering of reviews. These techniques are K means, mini batch K means, Affinity Propagation, and DBSCAN.

- **K-means :** The K-means clustering algorithm was first proposed by John Hartigan [150]. In this process, the initial cluster centers are first found out randomly and on the basic of that the process of clustering continues [151]. The steps for K-means clustering can be exaplined as follows:

  - A dataset D = $\{d_1$,d_2,d_3,.....$d_n\}$ consist of 'n' different data point or features.
  - In k-means, the number of clusters are defined before the processing starts. Here in this case two clusters are defined i.e., positive and negative cluster.
  - The squared Euclidean distances between the features and the centroid (cluster center) are found out. This value is known as clustering error and varies upon the center of cluster.
  - This error can be found out using following equation:

$$E(c_1, ..., c_m) = \sum_{i=1}^{N} \sum_{k=1}^{M} I(d_i \in c_k) \|d_i - c_k\| \tag{7.1}$$

    where:
    $E(c_1, c_2, ..., c_m)$ is the error found out for different cluster,

$$I(d_i) = \begin{cases} 1 & if\ D\ is\ positive \\ 0 & if\ D\ is\ negative \end{cases}$$

    $\|d_i - c_k\|$ indicates the distance between the features and the center.
  - Depending up on the distance of the data point form the centroid, the centroid is changed until the optimum result obtained where the data points make a cluster near centroid.

- **Mini Batch K-means:** Mini-Batch K-Means is modified form of K-Means Method. Its uses smaller subset to decrease the processing time and trying to increase optimize solution [152]. Each subset is randomly created in every iteration. To find the local solution of problem, mini batch reduces the computation. From the result obtained, it is observed that it is no way superior to any standard algorithm.

  The algorithm has basically two steps. In first step, different samples are selected randomly from the dataset to create mini batch. Those mini batch created are allocated to nearest centroid. In next step centroid gets updated. For each sample the above mentioned steps are repeated. For each subset of data in mini batch, centroid gets updated by average of sample data and all previous sampled data in that particular centroid. This process helps in decreasing the rate of change of centroid over time. All those steps are repeated till fixed number of iterations are reached.

The mini batch K means algorithm is an optimization problem to find out the set of clusters C, in order to minimize over a set of data X having an objective function as:

$$min \sum_{x \in X} \|f(C, x)\|^2 \tag{7.2}$$

where: $f(C, x)$ returns the nearest cluster center to x using Euclidean distance.

- **Affinity Propagation:** This algorithm finds the similarity between pair of input data point. Several messages are exchanged between data points until the best set of exemplars comes out. Here the exemplar refers to representative of each cluster [153]. The approach can be explained as follows:

  - The dataset D=$\{d_1$,d_2,...,$d_n\}$ are the 'n' different data elements or features where the exemplar can be represented as d_i.
  - 's' be the function that represents the similarity between two data points, where:
    $s(x_i, x_j) > s(x_i, x_k)$
    iff x_i is more similar to x_j than x_k.
  - The algorithm moves forward with updating the message passing steps, thus creating two different matrices i.e., ''Responsibility matrix'' and ''Availability matrix''. All these matrices are initially set to zero and then updated as the process continues.
  - The responsibility matrix R has values r(i, k) that quantifies as to how x_k serves as the exemplar for x_i , relative to other candidate exemplars for x_j. The matrix can be updated as follows:

    $$r(i, k) \leftarrow s(i, k) - \max_{k' \neq k} \{a(i, k') + s(i, k')\} \tag{7.3}$$

    where:
    $s(i, k)$ represents the similarities between the data item i and k.
  - The ''availability '' matrix A contains values a(i, k) that represents as to how ''appropriate'', it would be for x_i to pick x_k as its exemplar, taking into account of other points' preference for x_k as an exemplar. The matrix can be updated as follows:

    $$a(i, k) \leftarrow min(0, r(k, k) + \sum_{i' \notin \{i,k\}} MAX(0, r(i^{'}, k) \quad if\ i \neq k \tag{7.4}$$

    $$a(k, k) \leftarrow \sum_{i \neq k} MAX(0, r(i', k)) \tag{7.5}$$

where:
$r(k,k)$ is the responsibility of cluster k for data point k

- **DBSCAN:** The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) technique was first proposed by Ester *et al.* [154]. This technique is used to discover the clusters and to reduce the noise between the data items. The steps carried out to obtain the final clustering using DBSCAN is as follows:
  - The process starts with any arbitrary data point d_i and finds out all other data points reachable from d_i.
  - If the point d_i is a core point then it can be a cluster center and if it is a border point, there is no point reachable from d_i. Thus, DBSCAN moves to any other point.
  - After finding out the random point d_i, a maximum distance m_i is set i.e., if the distance between two data points are less than the maximum point, then they are considered in a cluster.
  - Depending upon the number of data item in the cluster and in order to avoid any overlapping of data point between the cluster, the process is repeated with changing m_i until the non-overlapping clusters are obtained.

7.3.5 Evaluation Parameters used

The performance of the unsupervised MLTs are often checked in order to have a comparative view. The performance evaluation parameters can be discussed as follows:

- **Homogeneity:** The data point that belongs to single class needs to be assigned to single cluster in order to satisfy homogeneity criteria, i.e., they are expected to have zero entropy [155]. If the data elements are perfectly homogeneous the conditional probability H(C|K) is zero where 'C' represents the class and 'K' represents the number of cluster. But in case of non-homogeneous data element, the value of H(C|K) depend on size of the dataset and also the distribution of class. Thus, homogeneity value of 'h' can be calculated as follows:

$$h = \begin{cases} 1 & if \;\; H(C|K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & if \;\; H(C|K) \neq 0 \end{cases} \tag{7.6}$$

- **Completeness:** For a given classes, all data points need to be member of same cluster in order to satisfy the criteria of completeness. In order to check the completeness, the distribution of cluster with in a class is examined. If the result is perfectly complete, it means that all data points from different classes are skewed into single cluster mentioned in [155]. The degree of skewness is evaluated by the conditional entropy

of the proposed class distribution given the class of component data points represented by H(K|C). For a prefect complete case, H(K|C) equals to zero. Thus, completeness value of 'c' can be calculated as follows:

$$c = \begin{cases} 1 & if \;\; H(K|C) = 0 \\ 1 - \frac{H(K|C)}{H(C)} & if \;\; H(K|C) \neq 0 \end{cases} \tag{7.7}$$

- **V-measure:** V-Measure is the weighted harmonic mean of homogeneity and completeness. It evaluates how successfully criteria of completeness and homogeneity are fulfilled, as described by Rosenberg and Hirschberg [155]. It's a entropy based measurement and equivalent to F-measure used to evaluate the performance in supervised MLTs . It is calculated by

$$V_\beta = \frac{(1+\beta) \times h \times c}{(\beta \times h) + c} \tag{7.8}$$

where: h indicates homogeneity,
c indicates completeness,
The value of β is greater than one if completeness is weighted more than homogeneity else less than one if homogeneity is given more weight than completeness.

- **Adjusted Rand Index (ARI):** Rand index in clustering is the measurement of similarity of data cluster. Adjusted Rand Index is another form of Rand index. In rand index the value obtained lies between 0 and 1, but in case of adjusted rand index values can be negative in case when index value is less than expected index. Although its value is similar to accuracy, but it is only applicable when there is no class label on data [156].

Given a set S of v elements, and two cluster of these points, namely $X_1, X_2, X_3, ...X_n$ and $Y_1, Y_2, Y_3, ...Y_r$. The overlapping of between X and Y can be summarized in a contingency table 7.1, where each entry v_{ij} denotes the number of objects in common between x_i and y_j.

$$ARI = \frac{i - e_i}{max_i - e_i} \tag{7.9}$$

$$ARI = \frac{\sum_{ij} \binom{v_{ij}}{2} - \frac{\left[\sum_i \binom{p_i}{2} \sum_j \binom{q_j}{2}\right]}{\binom{v}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{q_j}{2}\right] - \frac{\left[\sum_i \binom{p_i}{2} \sum_j \binom{q_j}{2}\right]}{\binom{v}{2}}} \tag{7.10}$$

where:
i indicates Index,
e_i indicates expected Index,
max_i indicates Maximum index

Table 7.1: Contingency table

	Y_1	Y_2	Y_3	. . .	Y_r	sums
X_1	v_{11}	v_{12}	v_{13}	. . .	v_{1r}	p_1
X_2	v_{21}	v_{22}	v_{23}	. . .	v_{2r}	p_2
X_3	v_{31}	v_{32}	v_{33}	. . .	v_{3r}	p_3
. .	. .	. .	. .	. .	. .	. .
X_n	v_{n1}	v_{n2}	v_{n3}	. . .	v_{nr}	p_n
sums	q_1	q_2	q_3	. . .	q_r	

- **Silhouette Coefficient:** It represents the comparison of tightness and separation of cluster. It shows which data point lies inside the cluster and which data points lie somewhere in between clusters [157]. Mathematically Silhouette coefficient can be defined as

$$s(i) = \frac{b_i - a_i}{max(a_i, bi)} \tag{7.11}$$

or

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & if \;\; a(i) > b(i) \\ 0, & if \;\; a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & if \;\; a(i) > b(i) \end{cases} \tag{7.12}$$

where:
i indicates each data point,
$a(i)$ indicates average dissimilarity of data within a cluster,
$b(i)$ indicates lowest average dissimilarity of other cluster where i does not belong to it.
Thus -1$\leq s(i) \leq$ 1.

7.4 Proposed Approach

The Figure 7.1 explains the step wise procedure for the clustering of the Twitter tweets using unsupervised machine learning techniques.

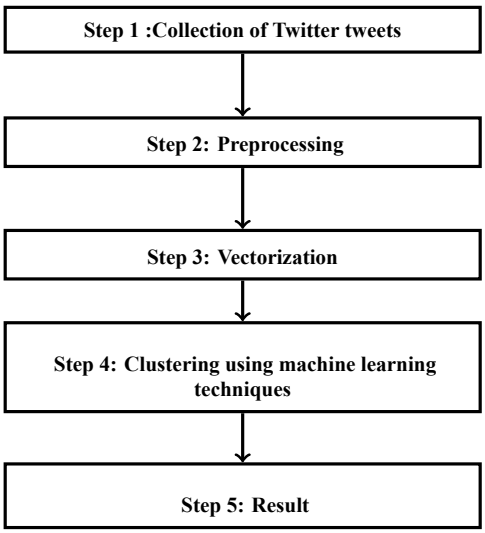

Figure 7.1: *Diagrammatic view of the proposed approach*

The detailed description of the steps are given below:

Step 1. The Twitter tweets are collected using TwitterAPI and R library called ‘‘twitterR’’. 24000 tweets based on #Indian politics are collected for the clustering.

Step 2. The tweets collected from Twitter contain some absurd information that need to be removed before the textual tweets transformed into numerical vectors. The absurd information are:

- Stop words: Stop words do not play any role in determining the sentiment, thus may be removed. The list of the stop words are collected from the site ‘‘ *http://norm.al/2009/04/14/list-of-english-stop-words/*’’ and then a list of those words being created and if the words appear in the text reviews then they are removed considered them as the stop words.
- Numeric and special character: In the text reviews, there are different numeric (1,2,...,5 etc.) and special characters (@, #, $,% etc.) present, which do not have any effect on the analysis; but they create confusion while conversion of text file to numeric vector.
- URL and HTML tags: These information also need to be removed as they do not play any role in finding out the sentiment.

After the absurd information are removed, the stemming process is carried out i.e., the process of getting the root word from any word. For example: the root word for reading is read. For the stemming purpose, PorterStemmer tool is used [158]. It is used to remove the common morphological and inflexional endings from words in English.

Step 3. After the removal of unwanted information, the next step is to convert the text reviews into numerical vector. Two different methods are often used for conversion of text data into numerical vectors which are CV and tf-idf. In this chapter, both CV and tf-idf methods are used for conversion of text data into numerical data.

Step 4. After the text data is converted into numerical vectors, they are given input to the unsupervised machine learning algorithms to obtain the clustering of reviews. The following algorithms are narrated in short as follows:

- K-Means: This algorithm is simple and fast for computation of clustering. In this algorithm initial cluster centers are assigned randomly which have a great impact on result formed. the distance of data points are calculated form the center and based on it the clustering is done.
- Mini batch K-Means: Its uses smaller subset to decrease the processing time and tries to increase optimize solution. In each step a random subset of total data is considered and with change in result the center changes to get optimum value.
- Affinity propagation: This algorithm finds the similarity between pair of input data point. Several messages are exchanged between data points until the best set of exemplars comes out. Here exemplar refers to representative of each cluster.
- DBSCAN: Clustering of data in DBSCAN algorithm is formed based on density of data. Clusters are separated between high density and low density.

Step 5. After the different machine learning algorithms are implemented, they are evaluated using different performance evaluation parameters. The result obtained is shown in following table 7.2 as below.

It can be observed from the Table 7.2, that the DBSCAN method shows the best result as compared to other three methods. It can also be noted that the values of homogeneity, completeness, V-measure and ARI are close to 1, where as the value of Silhouette coefficient is close the zero i.e., the parameters other than Silhouette coefficient need to be higher to show the better accuracy and the Silhouette coefficient value need to be low enough indicating the error rate.

The Figure 7.2 shows a comparative analysis of the performance evaluation parameter of proposed algorithms. This figure shows a graphical comparison of the values obtained after the clustering process i.e., the performance evaluation parameters. It can be observed that the homogeneity value varies in a range of 0.745 to 0.953, the completeness value varies in between 0.76 to 0.88, the V-measure varies in between 0.75 to 0.91, ARI otherwise known to be accuracy varies from 83 to 95, Finally silhouette value varies from 0.007 to 0.004.

Table 7.2: Performance evaluation after clustering

MLTs used	Evaluation Parameters				
	Homogeneity	Completeness	V-measure	Adjusted Rand Index	Silhouette Coefficient
K-means	0.745	0.764	0.754	0.834	0.007
Mini Batch K-means	0.626	0.675	0.650	0.704	0.006
Affinity Propagation	0.912	0.854	0.882	0.85	0.111
DBSCAN	0.953	0.883	0.917	0.95	0.004

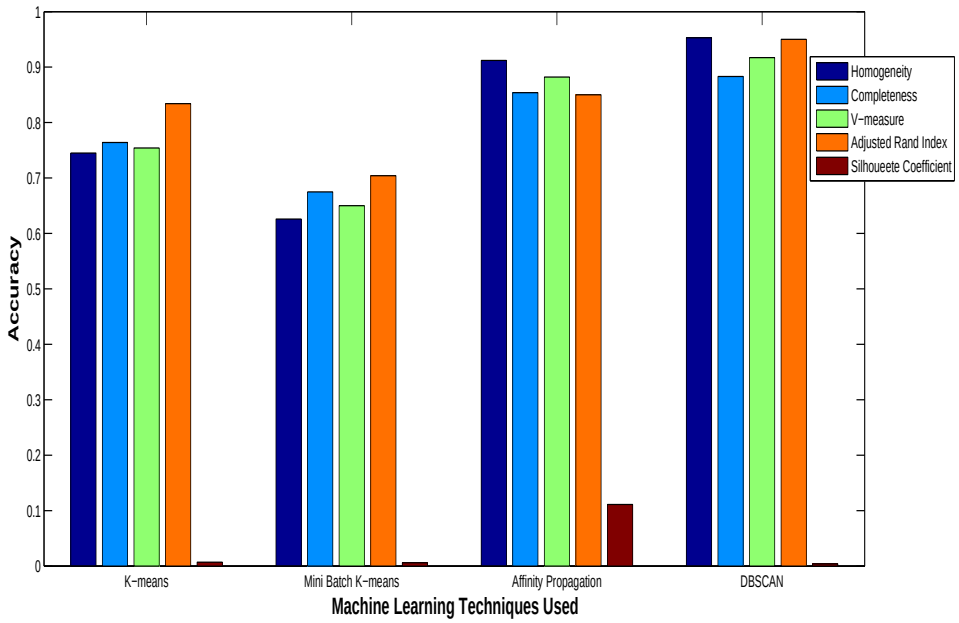

Figure 7.2: *Comparison of results obtained using proposed approach*

The DBSCAN method shows a better result as compared to other methods because in this method, the analysis is mainly based on the density or distribution of the data element. On the other hand, in the case of k-means and mini batch k means the analysis is based on the distance of the data points from the centroid which goes on changing until the optimum result is obtained. Thus, in these cases the result found out to be less accurate. Even in case of Affinity Propagation, where message transmission between the data points carried out and the comparison between the messages indicates the center and associated cluster. Thus, the DBSCAN method

shows better result in comparison with other methods as it works on distribution of the data points that helps in formation of clusters.

7.5 Performance Evaluation

The Table 7.3 compares result of the proposed method with the result obtained by different authors.

Table 7.3: Comparative analysis of obtained result with result of other authors

Methods	Authors					
	Li and Liu [147]	Balabantaray et al. [159]	Chaturvedi et al. [160]	Sureka and Punitha [161]	Scully et al. [152]	Proposed Approach
K-means	78.33	66.67	⊗	⊗	⊗	83.4
Mini Batch K-means	⊗	⊗	⊗	⊗	65.38	70.4
Affinity Propagation	⊗	⊗	75.06	⊗	⊗	85
DBSCAN	⊗	⊗	⊗	91.66	⊗	95.2

⊗ indicate that the algorithm is not considered by the author in their respective paper

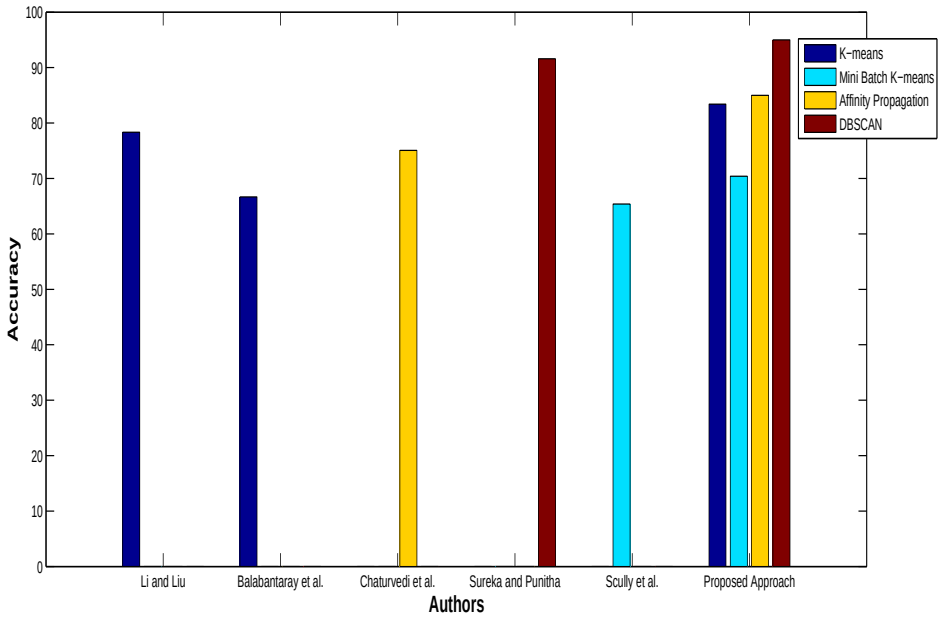

Figure 7.3: *Comparison of obtained results with result obtained by other authors*

From Table 7.3 and Figure 7.3, it can be viewed that Li and Liu [147] have obtained an accuracy for K-Means algorithm is 78.33% where as Balabantaray et al. [159]have found out an accuracy of 66.67%. But in proposed approach the accuracy achieved is 83.44%. In case of application of mini batch K means algorithm, Sculley et al. [152], have achieved an accuracy of 65.38%, But in proposed approach the accuracy achieved is 70.04%. In case of Affinity propagation, Chaturvedi et al., [160], have obtained an accuracy of 75.06% where as the proposed method shows a result of 85%. In case of application of algorithm DBSCAN algorithm, the accuracy obtained by Sureka and Punitha is 91.66 % but as per the proposed approach the accuracy is 95.20 %.

7.6 Summary

In this chapter, sentiment clustering is carried out using unsupervised machine learning techniques. Twitter tweets are collected using TwitterAPI and R library called ‘‘twitterR’’. 24000 tweets are collected from Twitter using #IndianPolitics. Four different machine learning techniques namely K means, mini batch K-means, Affinity Propagation, and DBSCAN are used for analysis. Different performance evaluation parameters are used to evaluate the performance of unsupervised MLTs. Among the four different MLTs, DBSCAN method shows the best result i.e., 95%.

Chapter 8

Conclusion

In this thesis, issues related to sentiment analysis are discussed with main focus on the classification of the reviews to obtain their sentiments in the form of positive or negative polarity. Comparison of the proposed result in this study with the results as reported in literature shows that the proposed result has performed better result in comparison with others. In this thesis, four different supervised classification approaches along with one unsupervised classification technique are discussed to improve the classification accuracy of the reviews.

Firstly, two different datasets IMDb [32] and Polarity [33] are considered for analysis as both have different approaches for analysis i.e., IMDb dataset has separate testing and training data whereas cross validation technique needs to be adopted in Polarity dataset as there is no separation between training and testing data. Four different machine learning techniques i.e., NB, SVM, RF and LDA have been implemented for classification of movie reviews. NB uses probabilistic Bayesian method for classification, SVM uses kernel based system for classification, RF uses an ensemble method, and LDA uses a discriminant analysis method for classification of reviews. Application of RF shows the best result among these four and also shows better results over those reported by different authors using both dataset. RF shows an accuracy of 88.8% in IMDb dataset and an accuracy of 95% in polarity dataset.

Secondly, n-gram approach is proposed, where the consecutive words are considered for analysis. In this thesis, unigram, bigram, trigram and their combination i.e., unigram + bigram, bigram + trigram, unigram + bigram + trigram are used for the classification. The IMDb dataset is considered for analysis. Again, four different MLTs are used for classification i.e., NB, SVM, ME, and SGD. It is found out that the accuracy of the classifier is better in case of lower level of n-gram i.e., for unigram but when the value of n increases, the corresponding accuracy value decreases. Among all the combination of MLTs and n-gram approach, SVM in combination with unigram + bigram + trigram approach has shown the best accuracy of 88.94%, which is better as compared to those of the result proposed by other authors.

Thirdly, an hybrid approach is proposed. The polarity dataset is considered for analysis. The text reviews are represented in the form of chromosome for the use of GA classifier. The GA finds the fit chromosomes from the set of chromosomes and provides them as

input to ANN for classification. Again, GA also uses different operators to select the best chromosomes and classify the reviews. Along with GA, ANN also performs the classification. During the classification process, the hidden nodes are kept on changing till the best possible accuracy is obtained. The GA classifier obtains an accuracy of 93% while NeuroGA i.e., combination of ANN and GA shows an accuracy of 96% which is found out to be better in comprison with results obtained in most of literatures.

fourthly, a feature selection approach is used for classification. The SVM is used to find out the sentiment values for each feature and based on these values, it selects the features with higher sentiment values both positive and negative polarity. These features are then given as input to ANN for classification. During the process the classification, the number of hidden nodes are kept on changing to obtain the best accuracy value. The feature selection process is carried out on both IMDb and polarity dataset, where 19729 features are selected from 159438 features and 3199 features are selected from 25579 features respectively. The accuracy obtained using this approach for IMDb and Polarity dataset are 95% and 96.4% respectively.

Finally, unsupervised approach is used to cluster the reviews collected from Twitter for analysis. Four different machine learning techniques are used to analyze the reviews and cluster them in to two different classes i.e., positive and negative. The tweets are collected using Twitter API. 42000 tweets related to politics are collected from Twitter for analysis. Then the tweet are preprocessed and finally clustered into two different cluster. Different performance evaluation parameters are used to evaluate the performance of the techniques. The four machine learning techniques i.e., K-means, mini batch k means, Affinity Propagation and DBSCAN have shown an accuracy of 83.4 %, 70.4 %, 85 % and 95% respectively.

8.1 Scope for Further Research

In this thesis the sentiment classification is carried out on supervised manner i.e., the dataset used for analysis is a labeled one. The training data is used for training and based on this, the testing of testing data is carried out. The polarity of the text are collected and compared to the original label to obtain the accuracy. Again, an attempt is made to perform unsupervised approach of sentiment analysis where there is no labeled data present for analysis. This study can be extended and the further research can be carried out in the following direction:

- It is observed that sometime, a small amount of labeled data is available on any topic and a large volume of data is unlabeled. In that case, semi-supervised approach may be adopted in which the unlabeled data is transformed to labeled data for further analysis.
- Again the source of data collection is also important. The dataset needs to be universally accessible which a large number of researchers often consider for their

analysis. In the present day scenario, the Twitter and Facebook tweets or comments are the important source of analysis of the reviews. These reviews need further analysis as discussed below for sentiment analysis.

  - Different reviews or comments contain symbols like (☺, ☹) which help in presenting the sentiment, but these images need application of special techniques for analysis.
  - In order to give stress on a word, it is observed that some persons often repeat the last character of the word a number of times such as ‘‘greatttt, Fineee’’. These words usually do not have a proper meaning; but they may be considered and further processed to identify sentiment,associated with the sentence as a whole.

- Deep learning approach may be used for the classification of sentiment reviews to check whether the deep learning approach shows an better accuracy result in compare to that of traditional methods.
- The performance of the proposed approaches are checked by using confusion matrix. But, in future, the performance must be check by rigorous statistical test as follows:
  - t test
  - ANOVA
  - Wilkinson RANK-SUM test
  - Wilkinson SIGN-RANK test
  - Sign test

  may be included to check the performance of the proposed system.

References

[1] B. Liu, ‘‘Sentiment analysis and opinion mining,’’ *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[2] J. A. Horrigan, ‘‘Online shopping,’’ *Pew Internet & American Life Project Report*, vol. 36, pp. 1–32, 2008.

[3] A. Lipsman, ‘‘Online consumer-generated reviews have significant impact on offline purchase behavior.” comscore,’’ *Inc. Industry Analysis*, pp. 2–28, 2007.

[4] T. Nasukawa and J. Yi, ‘‘Sentiment analysis: Capturing favorability using natural language processing,’’ in *Proceedings of the 2nd international conference on Knowledge capture, Sanibel Island, FL, USA*. ACM, 2003, pp. 70–77.

[5] K. Dave, S. Lawrence, and D. M. Pennock, ‘‘Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,’’ in *Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary*. ACM, 2003, pp. 519–528.

[6] C. Elkan, ‘‘Method and system for selecting documents by measuring document quality,’’ Apr. 3 2007, uS Patent 7,200,606.

[7] R. Feldman, ‘‘Techniques and applications for sentiment analysis,’’ *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[8] B. Pang, L. Lee, and S. Vaithyanathan, ‘‘Thumbs up?: sentiment classification using machine learning techniques,’’ in *Proceedings of the Association of Computational Linguistics conference on Empirical methods in natural language processing, Stroudsburg, PA, USA*, vol. 10. Association for Computational Linguistics, 2002, pp. 79–86.

[9] 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, Philadelphia, Pennsylvania*. Association for Computational Linguistics, 2002, pp. 417–424.

[10] J. M. Wiebe, R. F. Bruce, and T. P. O’Hara, ‘‘Development and use of a gold-standard data set for subjectivity classifications,’’ in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics,College Park, Maryland*. Association for Computational Linguistics, 1999, pp. 246–253.

[11] M. Hu and B. Liu, ‘‘Mining and summarizing customer reviews,’’ in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA*. ACM, 2004, pp. 168–177.

[12] B. Liu, ‘‘Sentiment analysis and subjectivity,’’ *Handbook of natural language processing*, vol. 2, pp. 627–666, 2010.

[13] N. Jindal and B. Liu, ‘‘Mining comparative sentences and relations,’’ in *Proceedings of American Association for Artificial Intelligence, Boston, Massachusetts*, vol. 22, 2006, pp. 1331–1336.

[14] J. Kamps, M. Marx, R. J. Mokken, M. d. Rijke *et al.*, ‘‘Using wordnet to measure semantic orientations of adjectives,’’ in *Proceedings of the Fourth International Conference on Language Resources and Evaluation, Centro Cultural de Belem, Lisbon, Portugal*. European Language Resources Association (ELRA), 2004, pp. 1115–1118.

[15] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.

[16] G. Mishne, N. S. Glance *et al.*, ''Predicting movie sales from blogger sentiment.'' in *in Proceedings of AAAI spring symposium: computational approaches to analyzing weblogs, Palo Alto, California*, 2006, pp. 155–158.

[17] E. Sadikov, A. Parameswaran, and P. Venetis, ''Blogs as predictors of movie success,'' in *in Proceeding of 3rd Int'l AAAI Conference on Weblogs and Social Media, San Jose, California*. Stanford InfoLab, 2009, pp. 1–5.

[18] F. Liu, B. Li, and Y. Liu, ''Finding opinionated blogs using statistical classifiers and lexical features.'' in *in Proceedings of 3rd International AAAI conference on Weblogs and Social Media, San Jose, California*, 2009, pp. 1–4.

[19] S. Asur and B. A. Huberman, ''Predicting the future with social media,'' in *in Proceeding of International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Omaha, Nebraska, USA*, vol. 1. IEEE, 2010, pp. 492–499.

[20] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, ''From tweets to polls: Linking text sentiment to public opinion time series.'' *4th International AAAI Conference on Weblogs and Social Media, George Washington University, Washington, DC*, vol. 11, pp. 122–129, 2010.

[21] A. Bermingham and A. F. Smeaton, ''On using twitter to monitor political sentiment and predict election results,'' in *5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011), Chiang Mai, Thailand*, 2011, pp. 2–11.

[22] N. A. Diakopoulos and D. A. Shamma, ''Characterizing debate performance via aggregated twitter sentiment,'' in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA*. ACM, 2010, pp. 1195–1198.

[23] E. T. K. Sang and J. Bos, ''Predicting the 2011 dutch senate election results with twitter,'' in *Proceedings of the Workshop on Semantic Analysis in Social Media, Avignon, France*. Association for Computational Linguistics, 2012, pp. 53–60.

[24] S. Das and M. Chen, ''Yahoo! for amazon: Extracting market sentiment from stock message boards,'' in *Proceedings of the Asia Pacific finance association annual conference (APFA), Bangkok, Thailand*, vol. 35, 2001, pp. 43–53.

[25] X. Zhang, H. Fuehres, and P. A. Gloor, ''Predicting stock market indicators through twitter "i hope it is not as bad as i fear",'' *Procedia-Social and Behavioral Sciences*, vol. 26, pp. 55–62, 2011.

[26] R. Bar-Haim, E. Dinur, R. Feldman, M. Fresko, and G. Goldstein, ''Identifying and following expert investors in stock microblogs,'' in *Proceedings of the Conference on Empirical Methods in Natural Language Processing,Edinburgh, UK*. Association for Computational Linguistics, 2011, pp. 1310–1319.

[27] J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li, and X. Deng, ''Exploiting topic based twitter sentiment for stock prediction.'' *Proceedings of the Association of Computational Lingustics-02 conference on Empirical methods in natural language processing, Sofia, Bulgaria*, vol. 2, pp. 24–29, 2013.

[28] M. S. Vohra and J. Teraiya, ''Applications and challenges for sentiment analysis: A survey,'' in *International Journal of Engineering Research and Technology*, vol. 2. ESRSA Publications, 2013, pp. 1–5.

[29] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, ''Supervised machine learning: A review of classification techniques,'' *Informatica*, vol. 31, pp. 249–268, 2007.

[30] T. Hastie, R. Tibshirani, and J. Friedman, ''Unsupervised learning,'' in *The elements of statistical learning*. Springer, 2009, pp. 485–585.

[31] M. F. A. Hady and F. Schwenker, ‘‘Semi-supervised learning,’’ in *Handbook on Neural Information Processing*. Springer, 2013, pp. 215–239.

[32] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, ‘‘Learning word vectors for sentiment analysis,’’ in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150.

[33] B. Pang and L. Lee, ‘‘A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,’’ in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Barcelona, Spain*. Association for Computational Linguistics, 2004, pp. 271–279.

[34] F. Salvetti, S. Lewis, and C. Reichenbach, ‘‘Automatic opinion polarity classification of movie,’’ *Colorado research in linguistics*, vol. 17, pp. 1–15, 2004.

[35] P. Beineke, T. Hastie, and S. Vaithyanathan, ‘‘The sentimental factor: Improving review classification via human-provided information,’’ in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Forum Convention Centre, Barcelona*. Association for Computational Linguistics, 2004, pp. 263–270.

[36] T. Mullen and N. Collier, ‘‘Sentiment analysis using support vector machines with diverse information sources.’’ in *Proceedings of EMNLP, Barcelona, Spain*, vol. 4, 2004, pp. 412–418.

[37] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, and W. Zuo, ‘‘Sentiment analysis of chinese documents: From sentence to document level,’’ *Journal of the American Society for Information Science and Technology*, vol. 60, no. 12, pp. 2474–2487, 2009.

[38] A. Yessenalina, Y. Yue, and C. Cardie, ‘‘Multi-level structured models for document-level sentiment classification,’’ in *in Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Massachusetts, USA*. Association for Computational Linguistics, 2010, pp. 1046–1056.

[39] M. Thomas, B. Pang, and L. Lee, ‘‘Get out the vote: Determining support or opposition from congressional floor-debate transcripts,’’ in *Proceedings of the 2006 conference on empirical methods in natural language processing, Sydney, Australia*. Association for Computational Linguistics, 2006, pp. 327–335.

[40] Z. Tu, Y. He, J. Foster, J. van Genabith, Q. Liu, and S. Lin, ‘‘Identifying high-impact sub-structures for convolution kernels in document-level sentiment classification,’’ in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, Jeju Island, Korea*. Association for Computational Linguistics, 2012, pp. 338–343.

[41] D. Bollegala, D. Weir, and J. Carroll, ‘‘Cross-domain sentiment classification using a sentiment sensitive thesaurus,’’ *IEEE transactions on knowledge and data engineering*, vol. 25, no. 8, pp. 1719–1731, 2013.

[42] R. Moraes, J. F. Valiati, and W. P. G. Neto, ‘‘Document-level sentiment classification: An empirical comparison between svm and ann,’’ *Expert Systems with Applications*, vol. 40, no. 2, pp. 621–633, 2013.

[43] D. Tang, B. Qin, and T. Liu, ‘‘Document modeling with gated recurrent neural network for sentiment classification,’’ in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal*, 2015, pp. 1422–1432.

[44] D. Zhang, H. Xu, Z. Su, and Y. Xu, ‘‘Chinese comments sentiment classification based on word2vec and svm perf,’’ *Expert Systems with Applications*, vol. 42, no. 4, pp. 1857–1863, 2015.

[45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ‘‘Efficient estimation of word representations in vector space,’’ *Computation and Language*, pp. 65–75, 2013.

[46] T. Joachims, ‘‘Training linear svms in linear time,’’ in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining,Philadelphia, PA, USA*. ACM, 2006, pp. 217–226.

[47] S. M. Liu and J.-H. Chen, ''A multi-label classification based approach for sentiment classification,'' *Expert Systems with Applications*, vol. 42, no. 3, pp. 1083–1093, 2015.

[48] M.-L. Zhang and Z.-H. Zhou, ''Ml-knn: A lazy learning approach to multi-label learning,'' *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[49] B. Luo, J. Zeng, and J. Duan, ''Emotion space model for classifying opinions in stock message board,'' *Expert Systems with Applications*, vol. 44, pp. 138–146, 2016.

[50] P. Ekman and W. V. Friesen, ''Constants across cultures in the face and emotion.'' *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[51] T. Niu, S. Zhu, L. Pang, and A. El Saddik, ''Sentiment analysis on multi-view social data,'' in *International Conference on Multimedia Modeling,Reykjavík, Iceland*. Springer, 2016, pp. 15–27.

[52] R. Xia, F. Xu, J. Yu, Y. Qi, and E. Cambria, ''Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis,'' *Information Processing & Management*, vol. 52, no. 1, pp. 36–45, 2016.

[53] J. Blitzer, M. Dredze, F. Pereira *et al.*, ''Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification,'' in *Proceedings of the Association of Computational Lingustics conference on Empirical methods in natural language processing*, vol. 7, 2007, pp. 440–447.

[54] D. Tang, ''Sentiment-specific representation learning for document-level sentiment analysis,'' in *in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining,Shanghai, China*. ACM, 2015, pp. 447–452.

[55] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, ''N-gram-based author profiles for authorship attribution,'' in *Proceedings of the conference pacific association for computational linguistics, Bali, Indonesia*, vol. 3, 2003, pp. 255–264.

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

[57] D. Bespalov, B. Bai, Y. Qi, and A. Shokoufandeh, ''Sentiment classification based on supervised latent n-gram analysis,'' in *Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, UK*. ACM, 2011, pp. 375–382.

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

[59] G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, and L. Chanona-Hernández, ''Syntactic n-grams as machine learning features for natural language processing,'' *Expert Systems with Applications*, vol. 41, no. 3, pp. 853–860, 2014.

[60] B. Agarwal, S. Poria, N. Mittal, A. Gelbukh, and A. Hussain, ''Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach,'' *Cognitive Computation*, vol. 7, no. 4, pp. 487–499, 2015.

[61] S. Foroozan, M. A. Murad, N. Sharef, and A. A. Latiff, ''Improving sentiment classification accuracy of financial news using n-gram approach and feature weighting methods,'' in *in Proceeding of 2nd International Conference on Information Science and Security (ICISS), Seoul, Korea*. IEEE, 2015, pp. 1–4.

[62] F. Aisopos, D. Tzannetos, J. Violos, and T. Varvarigou, ''Using n-gram graphs for sentiment analysis: An extended study on twitter,'' in *in Proceeding of Second International Conference on Big Data Computing Service and Applications (BigDataService), Oxford, UK*. IEEE, 2016, pp. 44–51.

[63] A. Abbasi, H. Chen, and A. Salem, ''Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums,'' *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 12, 2008.

[64] W. X. Zhao, J. Jiang, H. Yan, and X. Li, ''Jointly modeling aspects and opinions with a maxent-lda hybrid,'' in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Massachusetts, USA*. Association for Computational Linguistics, 2010, pp. 56–65.

[65] G. Ganu, N. Elhadad, and A. Marian, ''Beyond the stars: Improving rating predictions using review text content.'' in *WebDB*, vol. 9. Citeseer, 2009, pp. 1–6.

[66] Y. Wu, Q. Zhang, X. Huang, and L. Wu, ''Phrase dependency parsing for opinion mining,'' in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 2009, pp. 1533–1541.

[67] R. Feldman, B. Rosenfeld, R. Bar-Haim, and M. Fresko, ''The stock sonar—sentiment analysis of stocks based on a hybrid approach,'' in *Proceedings of Twenty-Third IAAI Conference, Hyatt Regency, San Francisco*, 2011, pp. 1642–1647.

[68] M. Govindarajan, ''Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm,'' *International Journal of Advanced Computer Research*, vol. 3, no. 4, pp. 139–149, 2013.

[69] A. S. H. Basari, B. Hussin, I. G. P. Ananta, and J. Zeniarja, ''Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization,'' *Procedia Engineering*, vol. 53, pp. 453–462, 2013.

[70] B. Agarwal and N. Mittal, ''Sentiment classification using rough set based hybrid feature selection,'' in *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis (WASSA'13), NAACL-HLT. Atlanta*, 2013, pp. 115–119.

[71] B. P. P. Filho, L. Avanço, T. A. S. Pardo, and M. d. G. V. Nunes, ''Nilc_usp: an improved hybrid system for sentiment analysis in twitter messages.'' in *in Proceeding of 8th International Workshop on Semantic Evaluation, Dublin, Ireland*. Association of Computational Linguistics Special Interest Group on the Lexicon-SIGLEX, 2014.

[72] F. H. Khan, S. Bashir, and U. Qamar, ''Tom: Twitter opinion mining framework using hybrid classification scheme,'' *Decision Support Systems*, vol. 57, pp. 245–257, 2014.

[73] B. Jagtap and V. Dhotre, ''Svm and hmm based hybrid approach of sentiment analysis for teacher feedback assessment,'' *International Journal of Emerging Trends of Technology in Computer Science (IJETCS)*, vol. 3, no. 3, pp. 229–232, 2014.

[74] K. Zhao and Y. Jin, ''A hybrid method for sentiment classification in chinese movie reviews based on sentiment labels,'' in *2015 International Conference on Asian Language Processing (IALP),Suzhou, China*. IEEE, 2015, pp. 86–89.

[75] V. Nandi and S. Agrawal, ''Political sentiment analysis using hybrid approach,'' *International Research Journal of Engineering and Technology (IRJET)*, vol. 03, 2016.

[76] M. Desai and M. A. Mehta, ''A hybrid classification algorithm to classify engineering students' problems and perks,'' *Computers and Society*, pp. 21–35, 2016.

[77] J. Neumann, C. Schnörr, and G. Steidl, ''Combined svm-based feature selection and classification,'' *Machine learning*, vol. 61, no. 1-3, pp. 129–150, 2005.

[78] C. Blake and C. J. Merz, ''{UCI} repository of machine learning databases,'' *Journal of machine learning research*, 1998.

[79] J. Weston, A. Elisseeff, B. Schölkopf, and M. Tipping, ''Use of the zero-norm with linear models and kernel methods,'' *Journal of machine learning research*, vol. 3, no. Mar, pp. 1439–1461, 2003.

[80] T. O'Keefe and I. Koprinska, ''Feature selection and weighting methods in sentiment analysis,'' in *Proceedings of the 14th Australasian document computing symposium, Sydney*. Citeseer, 2009, pp. 67–74.

[81] A. Esuli and F. Sebastiani, ''Sentiwordnet: A publicly available lexical resource for opinion mining,'' in *in Proceedings of Languare Resource and Evaluation, Genoa, Italy*, vol. 6. Citeseer, 2006, pp. 417–422.

[82] C. Nicholls and F. Song, ''Comparison of feature selection methods for sentiment analysis,'' in *Canadian Conference on Artificial Intelligence, ottawa, Ontario, Canada*. Springer, 2010, pp. 286–289.

[83] A. R. Webb, *Statistical pattern recognition*. John Wiley & Sons, 2003.

[84] S. Maldonado, R. Weber, and J. Basak, ''Simultaneous feature selection and classification using kernel-penalized support vector machines,'' *Information Sciences*, vol. 181, no. 1, pp. 115–128, 2011.

[85] A. Sharma and S. Dey, ''A comparative study of feature selection and machine learning techniques for sentiment analysis,'' in *in Proceedings of the 2012 ACM Research in Applied Computation Symposium, San Antonio, TX, USA*. ACM, 2012, pp. 1–7.

[86] A. Duric and F. Song, ''Feature selection for sentiment analysis based on content and syntax models,'' *Decision support systems*, vol. 53, no. 4, pp. 704–711, 2012.

[87] R. Xia, C. Zong, X. Hu, and E. Cambria, ''Feature ensemble plus sample selection: domain adaptation for sentiment classification,'' *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 10–18, 2013.

[88] O. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, ''A genetic algorithm-based feature selection,'' *British Journal of Mathematics & Computer Science*, vol. 4, no. 21, pp. 889–905, 2014.

[89] L. Zheng, H. Wang, and S. Gao, ''Sentimental feature selection for sentiment analysis of chinese online reviews,'' *International Journal of Machine Learning and Cybernetics*, pp. 1–10, 2015.

[90] B. Agarwal and N. Mittal, ''Prominent feature extraction for review analysis: an empirical study,'' *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 28, no. 3, pp. 485–498, 2016.

[91] A. K. Uysal, ''An improved global feature selection scheme for text classification,'' *Expert Systems with Applications*, vol. 43, pp. 82–92, 2016.

[92] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, ''A feature selection method based on improved fisher's discriminant ratio for text sentiment classification,'' *Expert Systems with Applications*, vol. 38, no. 7, pp. 8696–8702, 2011.

[93] H. Kanayama and T. Nasukawa, ''Fully automatic lexicon expansion for domain-oriented sentiment analysis,'' in *Proceedings of the 2006 conference on empirical methods in natural language processing, Sydney, Australia*. Association for Computational Linguistics, 2006, pp. 355–363.

[94] X. Wan, ''Using bilingual knowledge and ensemble techniques for unsupervised chinese sentiment analysis,'' in *Proceedings of the conference on empirical methods in natural language processing, Waikiki, Honolulu, Hawaii*. Association for Computational Linguistics, 2008, pp. 553–561.

[95] T. Zagibalov and J. Carroll, ''Automatic seed word selection for unsupervised sentiment classification of chinese text,'' in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Manchester, UK*. Association for Computational Linguistics, 2008, pp. 1073–1080.

[96] J. Rothfels and J. Tibshirani, ''Unsupervised sentiment classification of english movie reviews using automatic selection of positive and negative sentiment items,'' *CS224N-Final Project*, 2010.

[97] C. Lin, Y. He, and R. Everson, ''A comparative study of bayesian models for unsupervised sentiment detection,'' in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning, Vancouver, Canada*. Association for Computational Linguistics, 2010, pp. 144–152.

[98] G. Paltoglou and M. Thelwall, ‘‘Twitter, myspace, digg: Unsupervised sentiment analysis in social media,’’ *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 4, p. 66, 2012.

[99] M. Ghosh and A. Kar, ‘‘Unsupervised linguistic approach for sentiment classification from online reviews using sentiwordnet 3.0,’’ *Int J Eng Res Technol*, vol. 2, no. 9, 2013.

[100] X. Hu, J. Tang, H. Gao, and H. Liu, ‘‘Unsupervised sentiment analysis with emotional signals,’’ in *Proceedings of the 22nd international conference on World Wide Web,Rio de Janeiro, Brazil*. ACM, 2013, pp. 607–618.

[101] R. P. Abelson, ‘‘Whatever became of consistency theory?’’ *Personality and Social Psychology Bulletin*, 1983.

[102] D. A. Shamma, L. Kennedy, and E. F. Churchill, ‘‘Tweet the debates: understanding community annotation of uncollected sources,’’ in *Proceedings of the first SIGMM workshop on Social media,Beijing, China*. ACM, 2009, pp. 3–10.

[103] M. Fernández-Gavilanes, T. Álvarez-López, J. Juncal-Martínez, E. Costa-Montenegro, and F. J. González-Castaño, ‘‘Unsupervised method for sentiment analysis in online texts,’’ *Expert Systems with Applications*, vol. 58, pp. 57–75, 2016.

[104] R. Biagioni, ‘‘Unsupervised sentiment classification,’’ in *The SenticNet Sentiment Lexicon: Exploring Semantic Richness in Multi-Word Concepts*. Springer, 2016, pp. 33–43.

[105] A. B. Goldberg and X. Zhu, ‘‘Seeing stars when there aren’t many stars: graph-based semi-supervised learning for sentiment categorization,’’ in *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, Stroudsburg, PA, USA*. Association for Computational Linguistics, 2006, pp. 45–52.

[106] B. Pang and L. Lee, ‘‘Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,’’ in *Proceedings of the 43rd annual meeting on association for computational linguistics, Ann Arbor, Michigan*. Association for Computational Linguistics, 2005, pp. 115–124.

[107] V. Sindhwani and P. Melville, ‘‘Document-word co-regularization for semi-supervised sentiment analysis,’’ in *in proceeding of Eighth IEEE International Conference on Data Mining, Pisa, Italy*. IEEE, 2008, pp. 1025–1030.

[108] P. Melville, W. Gryc, and R. D. Lawrence, ‘‘Sentiment analysis of blogs by combining lexical knowledge with text classification,’’ in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France*. ACM, 2009, pp. 1275–1284.

[109] G. Lazarova and I. Koychev, ‘‘Semi-supervised multi-view sentiment analysis,’’ in *Computational Collective Intelligence*. Springer, 2015, pp. 181–190.

[110] D. Anand and D. Naorem, ‘‘Semi-supervised aspect based sentiment analysis for movies using review filtering,’’ *Procedia Computer Science*, vol. 84, pp. 86–93, 2016.

[111] A. M. Dai and Q. V. Le, ‘‘Semi-supervised sequence learning,’’ in *Advances in Neural Information Processing Systems*, 2015, pp. 3079–3087.

[112] R. Johnson and T. Zhang, ‘‘Semi-supervised convolutional neural networks for text categorization via region embedding,’’ in *Advances in neural information processing systems*, 2015, pp. 919–927.

[113] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer *et al.*, ‘‘Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia,’’ *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[114] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, ‘‘Rcv1: A new benchmark collection for text categorization research,’’ *Journal of machine learning research*, vol. 5, no. Apr, pp. 361–397, 2004.

[115] N. F. F. D. Silva, L. F. Coletta, and E. R. Hruschka, ‘‘A survey and comparative study of tweet sentiment analysis via semi-supervised learning,’’ *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, p. 15, 2016.

[116] H. Scudder, ‘‘Probability of error of some adaptive pattern-recognition machines,’’ *IEEE Transactions on Information Theory*, vol. 11, no. 3, pp. 363–371, 1965.

[117] A. Blum and T. Mitchell, ‘‘Combining labeled and unlabeled data with co-training,’’ in *Proceedings of the eleventh annual conference on Computational learning theory, Madison, WI, USA*. ACM, 1998, pp. 92–100.

[118] F. H. Khan, U. Qamar, and S. Bashir, ‘‘Swims: Semi-supervised subjective feature weighting and intelligent model selection for sentiment analysis,’’ *Knowledge-Based Systems*, vol. 100, pp. 97–111, 2016.

[119] Q. Wang, K. Zhang, Z. Chen, D. Wang, G. Jiang, and I. Marsic, ‘‘Enhancing semi-supervised learning through label-aware base kernels,’’ *Neurocomputing*, vol. 171, pp. 1335–1343, 2016.

[120] N. F. F. da Silva, L. F. Coletta, E. R. Hruschka, and E. R. Hruschka Jr, ‘‘Using unsupervised information to improve semi-supervised tweet sentiment classification,’’ *Information Sciences*, vol. 355, pp. 348–365, 2016.

[121] A. Acharya, E. R. Hruschka, J. Ghosh, and S. Acharyya, ‘‘C 3e: A framework for combining ensembles of classifiers and clusterers,’’ in *in Proceeding of International Workshop on Multiple Classifier Systems, Naples, Italy*. Springer, 2011, pp. 269–278.

[122] S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov, ‘‘Semeval-2014 task 9: Sentiment analysis in twitter,’’ in *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014), Dublin, Ireland*, 2014, pp. 73–80.

[123] J. HLTCOE, ‘‘Semeval-2013 task 2: Sentiment analysis in twitter,’’ *Atlanta, Georgia, USA*, vol. 312, 2013.

[124] T. Miyato, A. M. Dai, and I. Goodfellow, ‘‘Virtual adversarial training for semi-supervised text classification,’’ *arXiv preprint arXiv:1605.07725*, 2016.

[125] R. Garreta and G. Moncecchi, *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd, 2013.

[126] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, ‘‘Learning word vectors for sentiment analysis,’’ in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 142–150.

[127] P. Refaeilzadeh, L. Tang, and H. Liu, ‘‘Cross validation, encyclopedia of database systems (edbs),’’ *Arizona State University, Springer*, pp. 6–20, 2009.

[128] A. McCallum, K. Nigam *et al.*, ‘‘A comparison of event models for naive bayes text classification,’’ in *AAAI-98 workshop on learning for text categorization,Madison, Wisconsin*, vol. 752. Citeseer, 1998, pp. 41–48.

[129] C. W. Hsu, C. C. Chang, and C. J. Lin, ‘‘A practical guide to support vector classification,’’ pp. 12–24, 2003.

[130] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, ‘‘Random forest: a classification and regression tool for compound classification and qsar modeling,’’ *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[131] L. Breiman, ‘‘Random forests,’’ *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[132] M. Welling, ‘‘Fisher linear discriminant analysis,’’ *Department of Computer Science, University of Toronto*, vol. 3, pp. 1–4, 2005.

[133] S. Balakrishnama and A. Ganapathiraju, ‘‘Linear discriminant analysis-a brief tutorial,’’ *Institute for Signal and information Processing*, vol. 18, pp. 23–45, 1998.

[134] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, ''Sentiment analysis and classification based on textual reviews,'' in *in Proceeding of International Conference on Information Communication and Embedded Systems (ICICES), Chennai, India*. IEEE, 2013, pp. 271–276.

[135] A. Aue and M. Gamon, ''Customizing sentiment classifiers to new domains: A case study,'' in *Proceedings of recent advances in natural language processing (RANLP), Borovets, Bulgaria*, vol. 1, no. 3.1, 2005, pp. 1–7.

[136] J. Read, ''Using emoticons to reduce dependency in machine learning techniques for sentiment classification,'' in *Proceedings of the Association of Computaional Linguistics student research workshop, Ann Arbor, Michigan*. Association for Computational Linguistics, 2005, pp. 43–48.

[137] A. Kennedy and D. Inkpen, ''Sentiment classification of movie reviews using contextual valence shifters,'' *Computational intelligence*, vol. 22, no. 2, pp. 110–125, 2006.

[138] C. Whitelaw, N. Garg, and S. Argamon, ''Using appraisal groups for sentiment analysis,'' in *Proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany*. ACM, 2005, pp. 625–631.

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

[140] L. Bottou, ''Stochastic gradient descent tricks,'' in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 421–436.

[141] J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[142] D. Beasley, R. Martin, and D. Bull, ''An overview of genetic algorithms: Part 1. fundamentals,'' *University computing*, vol. 15, pp. 58–58, 1993.

[143] 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.

[144] S. Jiang, G. Pang, M. Wu, and L. Kuang, ''An improved k-nearest-neighbor algorithm for text categorization,'' *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503–1509, 2012.

[145] G. P. Zhang, ''Neural networks for classification: a survey,'' *IEEE Transactions onSystems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 30, no. 4, pp. 451–462, 2000.

[146] D. Reby, S. Lek, I. Dimopoulos, J. Joachim, J. Lauga, and S. Aulagnier, ''Artificial neural networks as a classification method in the behavioural sciences,'' *Behavioural processes*, vol. 40, no. 1, pp. 35–43, 1997.

[147] G. Li and F. Liu, ''Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions,'' *Applied intelligence*, vol. 40, no. 3, pp. 441–452, 2014.

[148] G. Salton and C. Buckley, ''Term-weighting approaches in automatic text retrieval,'' *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[149] S. H. Al-Harbi and V. J. Rayward-Smith, ''Adapting k-means for supervised clustering,'' *Applied Intelligence*, vol. 24, no. 3, pp. 219–226, 2006.

[150] J. A. Hartigan and M. A. Wong, ''Algorithm as 136: A k-means clustering algorithm,'' *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[151] A. Likas, N. Vlassis, and J. J. Verbeek, ''The global k-means clustering algorithm,'' *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[152] D. Sculley, ''Web-scale k-means clustering,'' in *Proceedings of the 19th international conference on World wide web, Raleigh, NC, USA*. ACM, 2010, pp. 1177–1178.

[153] B. J. Frey and D. Dueck, ‘‘Clustering by passing messages between data points,’’ *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[154] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, ‘‘A density-based algorithm for discovering clusters in large spatial databases with noise.’’ in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[155] A. Rosenberg and J. Hirschberg, ‘‘V-measure: A conditional entropy-based external cluster evaluation measure.’’ in *in proceeding of EMNLP-CoNLL, Prague, Czech Republic*, vol. 7, 2007, pp. 410–420.

[156] W. M. Rand, ‘‘Objective criteria for the evaluation of clustering methods,’’ *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[157] P. J. Rousseeuw, ‘‘Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,’’ *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[158] M. Porter, *New models in probabilistic information retrieval*. University of Cambridge, Computer Laboratory, 1980.

[159] R. C. Balabantaray, C. Sarma, and M. Jha, ‘‘Document clustering using k-means and k-medoids,’’ *International Journal of Knowledge Based Computer Systems*, vol. 1, pp. 12–25, 2015.

[160] A. Chaturvedi, K. Barse, and R. Mishra, ‘‘Affinity propagation based document clustering using suffix tree,’’ in *International Journal of Engineering Research and Technology*, vol. 3, no. 1 (January-2014). ESRSA Publications, 2014.

[161] V. Sureka and S. Punitha, ‘‘Approaches to ontology based algorithms for clustering text documents,’’ *Proceedings of International Journal of Computer Technology and Application*, vol. 3, pp. 1813–1817, 2012.

Dissemination

Internationally indexed journals [1]

1. **Abinash Tripathy** and Santanu Kumar Rath, (2015).Object Oriented Analysis using Natural Language Processing concepts: A Review. **International Journal of Information Processing (IJIP)**, Volume : 9, Issue: 3, pp. 28- 38. ISSN: 0973-8215.

2. **Abinash Tripathy**, Ankit Agrawal, and Santanu Kumar Rath, (2016). Classification of Sentiment Reviews using N-gram Machine Learning Approach. **Expert System with Application, Elsevier**, Volume:57, pp. 117-126. DOI: 10.1016/j.eswa.2016.03.028

3. **Abinash Tripathy**, and Santanu Kumar Rath, (2016) Classification of Sentiment of Reviews using Supervised Machine Learning Techniques **International Journal of Rough Sets and Data Analysis (IJRSDA), IGI Global**, Volume:4(1), pp. 56-74, DOI: 10.4018/IJRSDA.2017010104.

4. **Abinash Tripathy**, Abhishek Anand and Santanu Kumar Rath. Document level Sentiment Classification using Hybrid Machine Learning Approach, **Knowledge and Information Systems, Springer**, pp. 1–27, 2017. [Online]. Available: http://dx.doi.org/10.1007/s10115-017-1055-z

Conferences [1]

1. **Abinash Tripathy**, Santanu Kumar Rath, (2014). Application of Natural Language Processing in Object Oriented Software Development. *IEEE 4th International Conference on Recent Trends in Information Technology (ICRTIT 2014)*, Madras Institute of Technology, Chennai-600044, 10-12 April, 2014, DOI: 10.1109/ICRTIT.2014.6996121

2. **Abinash Tripathy**, Ankit Agrawal, Santanu Kumar Rath (2014). Requirement Analysis Using Natural Language Processing. *5th International Conference on Advances in Computer Engineering (ACE2015)*, Kochi, Kerela, India, 26-27 December, 2014.

[1] Articles already published, in press, or formally accepted for publication.

3. **Abinash Tripathy**, Ankit Agrawal, Santanu Ku. Rath (2015) Classification of Sentiment of Reviews using Supervised Machine Learning Techniques, *3rd International Conference on Recent Trends in Computing (ICRTC)*, SRM University, Ghaziabad, India, 12-16 March 2015.

4. Nishant Kumar, **Abinash Tripathy**, Santanu Ku. Rath (2016) Sentiment Clustering of Movie Review Data using Unsupervised Machine Learning Algorithm, *CSIR sponsored 6th National Conference on Indian Language Computing (NCILC-2016)*, Cochin University Of Science And Technology (CUSAT), Kerala, 26-27 February 2016.

5. **Abinash Tripathy**, Nishant Kumar, Santanu Ku. Rath (2016) Neuro-Fuzzy Clustering Approach for Sentiment Analysis, *IEEE Sponsored International Conference On Innovations In Information, Embedded And Communication Systems (ICIIECS '16)*, Karpagam College of Engineering, Coimbatore, Tamilnadu, 17-18 March 2016.

Article under preparation [2]

1. **Abinash Tripathy**, Abhishek Anand and Santanu Kumar Rath. Document level Sentiment Analysis using Genetic Algorithm and NeuroGA *Applied Soft Computing, Elsevier*, 2016.

2. **Abinash Tripathy**, Abhishek Anand and Santanu Kumar Rath. Document level Sentiment Classification using Feature Selection Techniques, *Neurocomputing, Elsevier*, 2016.

[2] Articles under review, communicated, or to be communicated.

Resume

Abinash Tripathy is presently perusing Ph.D. in Department of Computer Science and Engineering. He is enrolled in doctoral research program from Department of Computer Science & Engineering at National Institute of Technology (NIT) Rourkela, India. He has completed M. Tech. in Computer Science & Engineering from Kalinga Institute of Industrial Technology (KIIT) University, Bhubaneswar, India in 2009. Prior to that, he has completed his Master is Computer Science (M. Sc.) from Utkal University, Bhubaneswar, Odisha, India in 2005. His research interests include Sentiment Analysis, Natural Language Processing, Software Engineering, and Service Oriented Architecture . He is a member of different professional bodies such as IACSIT, IAENG, UACEE and IRED.

E-mail: abi.tripathy@gmail.com
Phone: +919437124235

Index