Consistency and Improvement in Accuracy of Sentiment Data Classification Using Model Stacking

Thesis submitted to

The Superior College, Lahore,

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy in Computer Science

by

Muhammad Azam

Roll No. PDCS14201

Session: 2013–2016

Faculty of Computer Science and Information Technology

THE SUPERIOR COLLEGE LAHORE

Author's Declaration

I Muhammad Azam hereby state that my PhD thesis titled "**Consistency and Improvement in Accuracy of Sentiment Data Classification Using Model Stacking"** is my own work and has not been submitted previously by me for taking any degree from this University or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my PhD degree.

Name of Student: Muhammad Azam

Date: _____

The Superior College, Lahore

Plagiarism Undertaking

I solemnly declare that research work presented in the thesis titled **"Consistency and Improvement in Accuracy of Sentiment Data Classification Using Model Stacking"** is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and the complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and University

Towards plagiarism. Therefore, I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of Ph.D. Degree, the University reserves the rights to withdraw/revoke my Ph.D. degree and that HEC and the University has the right to publish my name on the HEC/University Website on which names of students are placed who submitted a plagiarized thesis.

Student/Author Signature: _____

Name: Muhammad Azam

The Superior College, Lahore

Certificate of Approval

This is to certify that the research work presented in this thesis, entitled **"Consistency and Improvement in Accuracy of Sentiment Data Classification Using Model Stacking"** was conducted by Muhammad Azam under the supervision of Dr. Engr. Brig. (R) Tanvir Ahmad.

No Part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the Faculty of Computer Science and Information Technology, The Superior College, Lahore in partial fulfillment of the requirements for the degree of Doctor of Philosophy in field of Computer Science and Information Technology at The Superior College, Lahore.

Student Name: **Muhammad Azam** Signature: _____

EXAMINATION JURY/BOARD APPROVAL:

Sr. No	Name	Role	Signature
1	Dr. Arfan Jaffar	Chair	
2	Dr. Engr. Brig.(R) Tanvir	Thesis Supervisor	
3	Dr. Usman Hashmi	Co-Supervisor	
4	Dr. Muhammad Hamid	External Examiner-1	
5	Dr. Muhammad Faheem Mushtaq	External Examiner-2	
6	Dr. Tehreem Masood	Internal Examiner	

Dedication

I dedicate my dissertation work to my family and many friends. A special feeling of gratitude to my loving parents (Late). Especially all my dedication to my late mother who was always praying for me in all my hard times, also my family which always stood beside me in difficult times. I also dedicate this dissertation to my many friends and colleagues who have supported me throughout the process. I will always appreciate all they have done.

I dedicate this work and give special thanks to one and only Prof. Dr. Ch. Abdul Rehman, the real motivator and special mentor. His motivation and support were all that was behind this research.

Acknowledgment

First, I would like to express my humble gratitude from the core of my heart to Almighty ALLAH, as it was impossible for me to complete my research thesis without HIS blessings.

Secondly it is my supervisor Prof. Dr. Tanvir Ahmed who always motivated me for my Ph.D. research and related research. I am certainly obliged thank him for his tolerance, inspiration, and vast knowledge. His supervision facilitated me during the whole period of research and writing of this dissertation. I could not have made-up having a better mentor and counsellor for my Ph.D. study.

In addition to my thesis supervisor, I would also like to thank my co-supervisor Dr. Usman Hashmi, who helped me in making mathematical models. And especially producing research papers.

I would also like to acknowledge Dr. Arfan Jaffar for his insightful comments and guidance, but also for the tough questions which forced me to broaden my research from numerous viewpoints.

And lastly very special acknowledgements to Mr. Fahad Sabah, Mr. Imran Saeed and Mr. Iftikhar Hussain, who were always there to help me to find solution to the problems related to programming. It was not possible for me to complete my write up without the help of these gentlemen. Without their precious support it would not be possible for me to conduct this research.

Muhammad Azam

Abstract

Improvement in accuracy has always been an area of interest for researchers. Classification of unstructured sentimental data is another challenge faced by data scientists. Data scientist uses different methodologies to improve accuracy of prediction. Different machine learning algorithms are considered better classifiers for different instances of data. Combination of classifiers called ensembles are also considered as better learning methodology. Different types of ensembles such as homogenous and heterogenous ensembles are widely in practice to classify different types of structured and unstructured data. Bagging and boosting are also considered as better classification techniques in homogenous classification. But no single method can be taken as consistent good performer on all kinds of data sets. A learning algorithm performing well on one data set may not be a good classifier for the other.

This research is an effort to overcome this problem. In this study, a comprehensive model is presented to improve the accuracy of two class sentimental data. The model presented in this dissertation consistently perform well, as it has been tested on three different data sets with different instance size. Amazon reviews, YELP reviews, and IMDB reviews are selected for this purpose. This model is applied on all mentioned data sets and note able improvement in accuracy is noticed as compared with individual classifiers and homogenous combiners.

Data set is learned by different heterogenous classifiers at base level. It must be taken in consideration that base level classifiers should be selected from different family of classifiers. Then output of these base level classifiers is combined to generate a data set which will act as an input for next level classifier. At meta level, the classifier with best results at base level is learned with the new data set formed. The model thus formed is then used to predict any query example provided.

Two different metrics are used to check improvement made by this model. One is accuracy measure, which tells us percentage of accurately predicted instances on test data. while other one is AUC-ROC measure, which tells the probability that how accurately a query sample will be predicted. The model presented in this study has shown improvement in both the measures.

Table of Contents

List of Figures

List of Tables

Chapter 1 : Introduction

1.1 Background

Plenty of data is stored on and being transferred between computers. One cannot really manage such a huge amount of data manually, even our capability to examine the data automatically lags the capability to store, retrieve and transfer it. The task of classification of text to manage the large-scale textual data effectively and get required information quickly has become a challenge. It is required to increase the performance of algorithms being used for classification.

The amount of data produced each day is enormous. Each day data of around 2.5 quintillion bytes is made, also 90% of this data was produced in last two year.

According a report generated by IDC (International Data Corporation) in May 2018, there are few mind-blowing stats about, how the data is increasing day by day. According to a report released by IDC on ever growing data sphere, the collective sum of world data will grow from 33ZB (Zeta Bytes) to 175 ZB by 2025 with annual increase of 61%. Imagine it (1ZB = 1 trillion GB). Report says that from the beginning of 2019, more data will be stored in core enterprise (cloud provider data centers) then in the all world's existing data endpoints (cell phones, pc's, other devices etc.). Since our research is concerned about text classification so here are few facts about text related data.

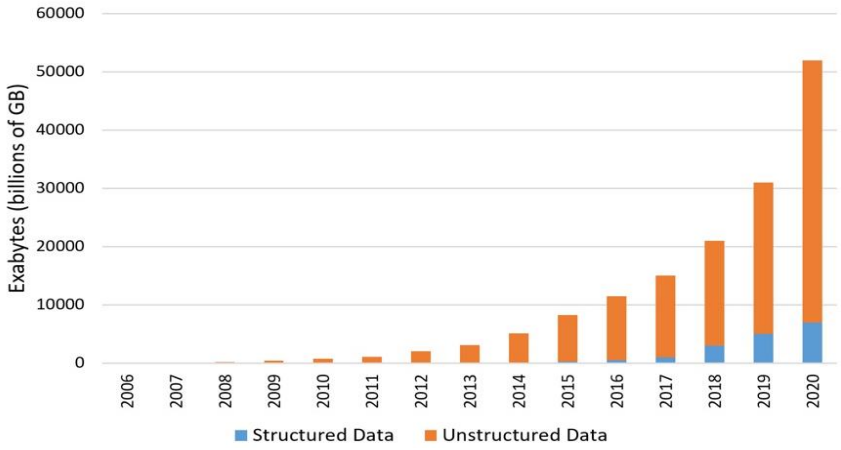

Figure 1.1 International Data Corporation Report on Structured vs. Unstructured Data

With such a huge amount of data (especially unstructured data), it is not possible to manage and process it with traditional available methods. Our capability to examine such a huge data lags far-off behind the ability to store, retrieve and transfer it.

The task of classification to manage a large-scale textual data effectively and get required information quickly has become a challenge. With increased need of data analysis on large scale data, there is a requirement for improving execution and scaling up of data mining and machine learning algorithms.

Text classification is the way toward arranging the content records dependent on words, phrases and word blends concerning set of predefined classifications. There are numerous applications of text classification, for example, fraud detection, research paper categorization, product reviews, mail directing, research paper categorization, email filtering, content arrangement, news categorization. Keywords are used for this purpose. Keywords are extracted from original text to classify the document. Keywords contains most substantial words that contain most important information about text data. These keywords are then processed to classify a document.

Comparative investigations of content classifiers have been done by different analysts. An investigation of customary content classifiers demonstrates that a few classifiers produce great results when there are only two classes. However, when applied on multiclass textual data, there performance is questionable. This encouraged the development of the multi-class classifiers, yet randomness was corresponding to the numbers of classes exist. This caused the preparation time of algorithm to rise quickly. A few algorithms cause a high-test time complexity which before long winds up impractical with huge sets. In the present day, speed is a significant part of documents classification alongside characterization exactness. Be that as it may, there is no reasonable victor on the two includes in a multiclass classification setup.

Researchers are not able to find a classifier or combination of classifiers that perform consistently well in all situations. The main issue which needs to be addressed that which combination or type of combination can be outperformer. Huge amount of text data is being transferred to servers every day. And this data can be isolated into a variety of subspaces. These subspaces generally have distinct features. Single classifier does not have the ability to consistently perform well on all these various subspaces of data. So, it is need of hour to a combination of classifiers that can produce better results on most of these subspaces.

This study introduces a new approach for the improvement in accuracy of classification systems. The core of the proposed methodology consists of improving the accuracy of classification by stacking, of two or more classification algorithms for classification. We will also try to use other ensemble techniques like bagging and boosting and compare the results with results of other

methods. The projected approach is capable to classify very big number of instances. In this study we offered, an improved classification system for classification of text data. We used three real time datasets which includes IMDB movies reviews, Amazon reviews for products and Yelp dataset.

1.2 Motivation

Ensemble learning algorithms are general techniques that enhance the prediction accuracy of classification models, for example, artificial neural networks, decision trees, Naïve Bayes, just as numerous different classifiers [1]. Outfit learning, in light of totaling the outcomes from various models, is a progressively refined methodology for expanding model precision when contrasted with the conventional routine with regards to parameter modification on a single model [2]. Ensemble method is a two-advance successive procedure comprising of a preparation stage where classification or predictive models are triggered from a preparation informational collection and a testing stage that assesses a totaled model against a holdout or concealed example. Notwithstanding the way that there has been general research related to solidifying estimators or guesses [3], gathering systems, as for arrangement calculations are new techniques. Thusly, it is basic to clarify the capability between gathering procedures and blunder approval strategies: troupe systems increase all around model precision while cross endorsement techniques increase the exactness of model mistake estimation [4].

Following figure describes different fields of machine learning. Heterogeneous classification is used in this study to improve accuracy of sentimental data.

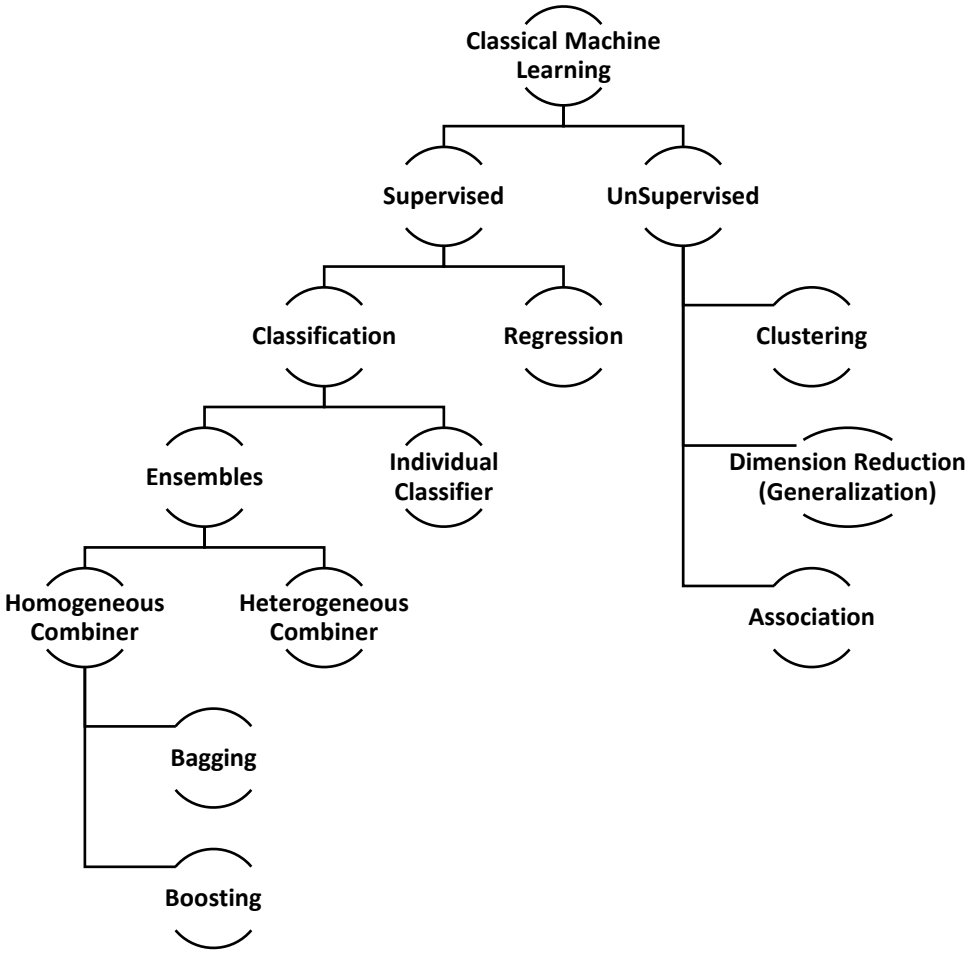

Figure 1.2 Machine Learning Algorithms

Since web data is endlessly growing so it is need of the hour that classifiers should be retrained on regular basis to keep up with increasingly changing data. Training time should also be kept in mind in this classification process. Memory requirement is another important factor which must be taken into consideration during this process. A better approach would be if time constraints and memory consumption should be taken into consideration. A classifier or combination of classifiers will be considered a better performer if it can produce better results consistently on different domains of text data while time/memory tradeoff should also be well thought out.

In the field of scientiometric research, systems of classification are necessary tools. A system of classification of scientific publications assigns journals or single publication to the research areas. Such a classification system can for be used to simplify literature search, to study the organization and underlying forces of scientific disciplines, or it can be used to facilitate scientiometric research evaluations. Over last two decades number of scientific publications has been extremely increased,

and this growth demands the effective organization and categorization of these documents. According to the data taken from scimago journal and country ranking number of publications is very high (more than one million) right from the beginning as depicted in Figure 1.3. Number of publications have been increased every successive year and it has been raised up to 3.5 million in recent years which is large digit.

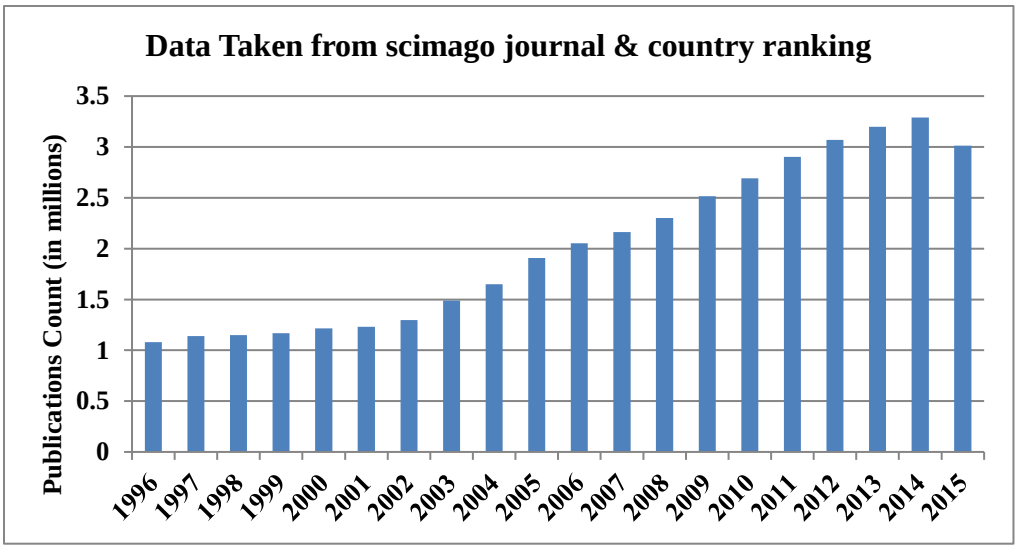

Figure 1.3 World's Scientific Publications from 1996 to 2015

While searching relevant publications, scientists must muddle through these exponentially growing numbers of research articles. Classification techniques can help scientists with the task as previously people used the algorithms for the recommendation and classification of music, movies and products. However, little is known about the performance of these algorithms with dataset containing scientific publications like SCOPUS. Additionally, only the archive of medicine comprises more than twenty-five thousand vocabulary terms to represent research areas. Whereas, there are around four hundred categories and sub-categories of music moreover, scientists may hide their findings awaiting full and final publication which means that completely novel areas might appear immediately. Scientific literature is not similar as other fields and hence particular requirements are required to be addressed.

1.3 Problem Statement

The efficient text processing and analyzing method is an important need that has grown due to immense creation of different kinds of digital resources which are taken from everywhere in structured or unstructured form. The objective of text processing in machine learning, text mining,

and natural language processing, etc. is to classify the documents or to categorize text in appropriate classes. It is important for classification algorithms to be efficient and give more accurate results. The current systems certainly have the potential to provide the required information; however, a tool or algorithm which can automate and outperform current systems at the level of individual publication is not yet available. The challenge is to design a classification process or an algorithm which can have consistently improved accuracy on all data sets and different instance sizes of same data sets than present algorithms.

The purpose of this work is to present improved classification model using a hybrid model in R to classify or categorize the textual data at the level of individual publication.

1.4 Objectives

This dissertation discusses four key areas that previously has not been explored.

- **Data representation.** Since data is obtained from internet. This sentimental data needs to be represented in proper form, so that it can be used in any programming tool. So, my first objective is to covert this informal data to formal form using a number of preprocessing steps.
- **Comparison of heterogenous base level classifiers.** A comprehensive comparison of performance of base level classifiers from different family of classifiers is next objective of this study. To find most efficient individual classifier for large text data is also objective at this level.
- **Performance of homogenous ensembles.** Another objective of this dissertation is to check whether homogenous combiners perform better than individual classifiers.
- **Develop a novel hybrid machine learning system.** Finally, designing an model to improve the accuracy and address consistency issue in accuracy improvement for classification and categorization of text data is last objective. A heterogenous ensemble approach is supposed to work to develop hybrid machine learning system.

1.5 Contributions:

Numerous advancements have been made in the field of text classification. Ensembles techniques are used to classify the text more efficiently. Major contributions of this thesis in the field of text classification are described below.

- Study and comparison of number of preprocessing steps and different samples sizes which affect the accuracy for classification of the text with individual classifiers.
- An efficient method is proposed to identify the ensembles/combination of classifiers for text classification.
- Solution for the classification of text problem using homogeneous classification techniques.
- Solution for the classification of text problem using heterogeneous classification techniques.

1.6 Structure of the Dissertation

This thesis consists of seven chapters. In the first chapter, an overview of sentiment analysis and data sources is provided. The problem statement, contributions and structure of thesis are also included in this chapter. In Chapter 2, an overview of related research for improvement in classification algorithms is presented, considering both the state of art methodologies and distinctive contributions to the subjected field. In Chapter 3, the datasets and pre-conditions I applied for processing data sets to be prepared for the analysis are discussed. In Chapter 4, different base classifiers from different family of classifications are used to train and test on their different data sets of same kind. Their results are compared and provided reasons which classifier is performing well and which is not.

In Chapter 5, homogeneous learning classifiers are discussed in detail. For bagging Rf and for boosting GBM classifiers are used. Both homogenous classifiers are trained and tested on different all three datasets and their results are compared with base classifiers.

In Chapter 6, a new variant, Heterogeneous Classifier, is introduced which improves Stacking performance consistently on all three datasets. Further, it reduces computational cost and resolves a important issues found during our experiments with base classifiers and homogenous classifiers.

Respective results in Chapters 4, 5 and 6 are presented from an alternative pattern to compare results of classifiers. I explored the hypothesis that heterogeneous ensemble is more consistant and stable than other ensemble learning schemes, i.e. both the metrics used i.e. accuracy measure and ROC curve shows improvement as compared to individual classifiers and homogenous ensembles. Finally, in Chapter 7, I concluded the thesis with a summary of our main findings, an overall

conclusion and a viewpoint on future endeavors. The limitations of the research are also presented in this chapter.

Chapter 2 : Historical Background

For many years, researchers related to machine learning conducted number of studies for the improvement of text classifiers with the combination of some classifiers created from more than one algorithm. Number of researchers worked on technique like; bagging and boosting. These techniques are based on creating groups of homogeneous classifiers, whereas, Model Stacking is another technique which is based on combining the heterogeneous classifiers. The researchers, who have addressed the Model Stacking problems, showed that there were serious issues while choosing base learning algorithms for creating meta-classifiers which are the members of the group. Number of studies on model stacking proposed manual selection of the suitable grouping of base learning algorithms. But some other studies proposed automatic systems to get good Stacking configurations.

2.1 Sentiment Analysis

People have different opinions about different product or services. Their expression can be different even about same product or service. It can be positive, negative or neutral. And there can be more classes for particular service or product. So, analyzing a sentiment can be defined as a process of determining whether an opinion is positive, negative or neutral. Following figure illustrate the working cycle of sentiment analysis.

Figure 2.1 Working Cycle of Sentiment Analysis

Formal definition of sentiment analysis can be “Sentiment analysis is the field of study that analyses people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards

entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" [5]

Liu defines opinions as "subjective expressions that describe people's sentiment, appraisals or feelings towards entities, events and their properties" [6]

2.2 Approaches of sentiment Analysis:

Primarily there are three method used for sentiment analysis i.e. machine learning based methods, lexical based methods and linguistic analysis.

2.2.1 Machine learning based methods:

Sentiment analysis based on machine learning methods are based on models that classify the textual data into different categories. Different machine learning classifiers or combination of classifiers are used to build models to categorize sentimental data. Models are used to train historical data and then tested on test data to find accuracy of model. If required accuracy is achieved then this model is adopted. These methods are discussed in details later in this section.

2.2.2 Lexical based methods:

Lexical based approach is most popular approach to analyze unsupervised text data. In lexical based methods a predefined list of words is used and each word is associated to a specific sentiment. A lexicon can be considered as a dictionary e.g. we can say that set of medical terms is a lexicon. And every document or piece of text can be considered as a lexicon. So, we can define lexicon as a dictionary or vocabulary of language [7].

Khin used lexicon-based approach to analyze student feedback in his study. He created a database of English sentiment words and uses it to analyze positive, negative and neutral feedback of students. For this purpose, they extracted sentiment words from students' feedback [8].

Another research was made about understanding the Vietnamese language and providing answers to a question asked. Kiet and his companions proposed a lexical based machine reading comprehension (MRC) technique exploits semantic similarity measures and external knowledge sources to analyze questions and extract answers from the given text. They found out that their proposed model achieved more than 5% accuracy as compared to best base level model [9].

Malicious advertisements have become a major issue of online advertising. Xhang, in his paper has tried to address this issue. He presented a lightweight online advertising classification system using lexical based features as an alternate solution. He used three different URL sources to generate three different scenarios. Then a set of lexical based features is selected for training and testing purpose. Results shows 97% of accuracy in some cases. Depending upon various URL

sources, they created three different URL sources. Then from previous, they selected a lexical based feature set. This set is used for training and testing purpose. Then this set is used for proposed model. Result shows that accuracy is about 97% in some cases. [10].

2.2.3 Linguistic analysis:

It is really an uphill task for computer to figure out what humans have written theoretically. We can define Linguistic Analysis as performing natural language processing (NLP) to understand and process the text written by human for further analysis. A variety of language can be used to describe a particular situation. To figure out the real meaning by natural language processing is linguistic analysis.

Cyber-attack is considered as one of major problem these days. Attackers can intrude enterprise networks and can damage important information. In his study, collected many lexical features and combined with blacklisted URLs to improve performance. They were able to get an acceptable 0.84 F-1 score [11].

Aditya Joshi1 and his colleagues worked Hindi-English code mixed data-set for sentiment analysis. They proposed learning sub-word level representations in LSTM (Subword-LSTM) architecture. They were able to attain 4-5 better accuracy than traditional methods [12].

2.3 Levels of Analysis:

One can work at different levels of sentiment analysis. Following are four levels of depth of sentiment analysis.

1. Document level.
2. Sentence level.
3. Phrase level.
4. User level.

2.3.1 Document level.

Whole document is targeted at document level sentiment analysis and an overall sentiment is assigned to it, assuming the whole document a single entity. Since there can be more than one opinion in whole document so, this assumption can be considered unrealistic. However, there are a number of cases like reviews, where end result of a complete document is single opinion. Another example can financial news, where news carries negative or positive sentiments. So, the text data

like movie reviews, product reviews, services reviews, financial news etc. are some examples of document level sentiment textual data.

2.3.2 Sentence level:

The existing work on Sentence-level sentiment analysis is focused on recognizing the polarity of a sentence (e.g. positive, neutral, negative), according to information learned from textual data.

A document contains a number of sentences. An assumption is taken into account that each sentence carries a single sentiment. which is certainly not the case. One cannot say that all the sentences are subjective. In fact, a lot of sentences are not subjective. i.e. these sentences do not carry any sentiment. So, first phase in these types of analysis must be to distinguish between subjective and non-subjective sentences. As non-subjective carry no information for any classifier. Subjective sentences are those which carry a sentiment or opinion.

Usually polarity of opinion consists of negative, positive or neutral sentiments. Different techniques are being used for analysis of sentence level sentiment analysis. Following are to name a few.

1. Corpus based techniques.
2. Dictionary based techniques.
3. Bootstrapping approach

Corpus-based procedures attempt to discover co-event examples of words to decide their sentiments

Dictionary-based methods use equivalent words, antonyms and chains of command in WordNet (or different vocabularies with notion data) to decide word sentiments.

Bootstrapping is another methodology. The thought is to utilize the yield of an accessible starting classifier to which a managed learning calculation might be applied.

Results from different studies shows that not a single classifier yields best results for different data sets. In fact, a combination of classifiers is better approach for sentence level sentiment analysis.

2.3.3 Phrase level:

Phrase or word level analysis researches the extremity of writings on a finer level: the phrase level. This includes separating polar phrases and afterward defining their sentiment. In finance, word

level sentiment is utilized to distinguish polar words and measure their connections to different factors like firm profit or stock costs.

As of late, numerous statistical models have been worked for a more profound analysis of item surveys, that is, mining the clients conclusions about certain item includes. This is usually alluded to as perspective level sentiment analysis. It is the way toward separating pertinent parts of the assessed item and deciding the sentiment of the relating feeling about them. The presumption here is that all sentiments are commonly aimed at a specific subject/object which nothing unless there are other options casing's objectives precisely and reliably. For instance, in film surveys, extricated angles could be: music, entertainers, or lights. At the point when the clients are expounding on a film, they express their conclusions about these angles like what they think about the entertainers or the music decision.

2.3.4 User level:

Notwithstanding the previous levels of analysis, some studies do the analysis on a user level that looks into users' networks and predicts user's sentiment based on the sentiment of neighboring users.

2.4 Sentiment Classification using machine learning approaches:

Classification refers to accurately predict the class of each query sample in the data. The historical data set is divided into training and test sets. A classifier is trained on training data and a model is formed. This model is then tested on test data set and accuracy percentage is found. If this model is adopted on the basis of accuracy achieved, then this model is applied on any data sample provided and required class is predicted.

1Fermi et. al. in their study presented a method for improvement openly planned to do working on Twitter data through reflection of their arrangement, length and semantic, an example of response investigation. Moving towards used dataset is just expandable to more languages and able sufficiently to development the tweets in actual instance. They presented that by using the preparation methods generated with the method defined in their study is be able to elevate correctness of categorization, despite of this field and sharing of test sets. [13]

Number of machine learning algorithms exist; Chris Thornton et. al. in their research proposed, that considering each algorithm's hyper parameters, there is an amazingly large number of potential alternatives. Going further than earlier work that attempts the issues of parameters selection for

learning algorithms separately, they considered the issue of simultaneously choosing a learning algorithm and setting its hyper parameters. They showed that this issue can be solved by completely automated method, leveraging modern improvements in Bayesian optimization. Precisely, they considered a number of feature selection techniques and all techniques of classification implemented in standard distribution of WEKA', spanning 2 ensemble methods, 10 meta-techniques, 27 base classifiers, and hyper parameter settings for every classifier. They showed that classification performance often much better than using standard selection and hyper parameter optimization methods on around 21 popular datasets in use from the UCI repository, variants of the MNIST dataset, the KDD Cup 09 and CIFAR-10. They expected that their method will help even unfamiliar users to more efficiently categorize the learning algorithms and hyper parameter settings suitable to their applications, and therefore to attain better performance [14] .

2.4.1 Sentiment Classification using individual machine learning classifiers:

Individual classification method does not outperform other classification techniques in all situations. Researchers tried to resolve the problem of classification by using numerous classification techniques and inspect their performance for classification purposes. Based on this performance, enhanced classification techniques could be adopted and techniques with bad performance might be evaded. But which individual classification technique is at top to determine the label of new instances is still vague, particularly when several methods give comparable classification performance? Krzyśkoet. al. presented numerous approaches, which combine many classification techniques to determine the label or a class of new instances. They showed that the best technique for combining classifiers is nonparametric regression [15].

Leevyet. al. addressed class imbalance problem in the dataset(s) They showed that class imbalance can dramatically reduce the accuracy of classifiers, due to the prediction biasness for the majority class. Leevyet. al. provided a survey of previous studies being one in the last eight years, with emphasis on class imbalance; having a class ratio between 100:1 and 10,000:1 (majority-to-minority). They covered two techniques which consist of Data-Level for example, data sampling and Algorithm-Level for example, cost-sensitive and ensemble techniques. They show that data sampling techniques are prevalent in dealing with class imbalances, and random oversampling techniques mostly show better results. At the algorithm level, there are some good performers. However, published researches have shown unpredictable and inconsistent results in addition to the limited range of methods evaluated, specifying that there is a need for more widespread

comparative studies. [15]

Junchi Bin et. al. worked on deep learning which is attracting substantial attention because of its computational capabilities in handling images as well as natural language processing. Increasingly real estate agents offer online intelligent systems to help their customers through the inquiry of directed properties and decision in advance of any transaction. The assessment of property is one of the greatest apprehensions for customers. In assessment procedure, approximation of the price of a house does not only depend on its characteristics but also influenced by its neighborhoods. Therefore, real estate assessment can be enriched if it comprises the feature of peer-dependence. Junchi Bin et. al. in their paper, proposed a valuation model based on peer-dependence (PDVM), this model is adeptin transforming the problem of peer-dependence-based valuation to an arrangement calculation issue. In their proposed idea, they developed a technique, "K-nearest similar house sampling (KNSHS)", which generates the arrangements from the to-be-value house and neighboring houses. Next, a "bidirectional long-term short-term memory (B-LSTM)" layer extracts the sequence type. Finally, the "Fully Connected (FC)" layer infers house prices based on the features. Their investigational results show that their proposed model overtakes other high-tech machine learning methods used for real estate valuation. [16]

MeteOzayaet. al. proposed a two-layer fusion method, known as Fuzzy Stacked Generalization (FSG) that established a hierarchal design based on remote learning. At the first stage of this technique, fuzzy k-NN learners accept completely dissimilar sets of feature these features are mined out from the equivalent dataset to achieve several observations of dataset. At the second-stage or meta-layer, a combination is created by accumulating the decisions of all the primary stage classifiers and afterward, fuzzy k-NN learner is prepared inside combination through diminishing the dissimilarity among the enormous example and N-test categorization mistake. To calculate the level of joint effort amongst the base-level learners and the variety of features, a new measure is introduced which is known as, shareability. This new measure has been laid out in light of properly classified samples by no less than one among the base-level learners in Fuzzy Stacked Generalization. They experimentally established that their system i.e. FSG performs better than the well-known combination methods once the shareability is sufficiently large such that large portion of the examples appropriately categorized by no less than one among the base-level classifiers. The tests executed on various artificial and genuine standard datasets demonstrated that the outcomes of FSG are improved as compared to the other combination techniques [17]

Rocchio is a prototypical way of doing the text classification in data mining and information retrieval and “Naïve-Bayes” and k-NN are the two mechanism for the categorization of text. Behzad Moshiriet. al. suggested advance way of text codification on the thrust of three methods and classifiers combination techniques. That is a pilot technique. In this technique documentation is categorized as vectors and each element of the vector is joint along certain term. For merging the classifiers, voting methodology, conclusion templates and OWA operator activity are offered. Their investigational outcomes reveal that the gain on decline the inaccuracy in categorization to 15% while they utilized instruction facts from 20 datasets of newsgroups [18].

C. Karthika et. al. projected a different text classifier through the grouping of nearest neighbor technique and “Support Vector Machines (SVM)”. Point of this learning recommended “SVM-NN” way is for reducing the result of boundaries in the precision of categorization. At preparation stage, the Support Vector Machines is functional to reduce the guidance examples for every class to their “support vectors (SVs)”. The Support Vectors from dissimilar categories are then used for the preparation facts of end-to-end surroundings in which the expanse purpose or likeness actions is used to evaluate which grouping does the trying data fits. The process also compacts interval utilization [19].

Some other scientists driven for the development of order calculations utilized “Genetic Algorithms (GAs)” (Goldberg, 1988) which join presence of fitting string structures using an organized however randomized data interchange to frame a hunt calculation. The subjected calculations have been utilized in characterization methods, AI and information mining applications [20] [21]. Hereditary Algorithms have additionally been utilized in enhancing some other learning strategies, as neural systems [22].

Nikolaos Mastrogiannis et. al. proposed “a method for improving the accuracy of data mining classification algorithms.” In this study they introduced a technique called; “Classification through ELECTRE and Data Mining (CL.E.D.M)”, which employs features of procedural outline of the ELECTRE I out ranking technique, with a goal of refining the accuracy of prevailing classification algorithms in data mining. Especially, the technique selects the unsurpassed decision rules taken from the training procedure of the data mining classifiers, and then it allocates the categories that related to these rules, to the instances that must be classified. In their study Nikolaos Mastrogiannis et. al. used three commonly used data mining algorithms for classification and tested them in five commonly used datasets to verify the strength of their proposed method [23].

Harshita Patel and G.S. Thakur, proposed a "Hybrid Weighted Nearest Neighbor Approach to Mine Imbalanced Data". A substantial research has been made during last few years on the classification of imbalanced datasets. Since ever growing huge text data can be distributed into a variety of classes, so traditional classification methods are not suitable for better performance. Therefore, there is need to modify traditional classification methods to deal such datasets. Authors, in this paper tried to overcome this problem by proposing a hybrid nearest neighbor learning methodology for mining imbalanced datasets. They used different K values for different sizes of datasets. They assumed large K value for large dataset and small K value for small dataset with weighted concept. The concept of different K values increased the performance of KNN weighted approach [24].

Rita Mandal and Shweta Yadav proposed "an Improved Intrusion System Design using Hybrid Classification Technique". The authors found out some serious issues with existing IDS approaches. Some of these needs huge data for processing. So, processing these large datasets requires more time for processing and also need more memory resources. As a result, it affects performance of classifier. In order to overcome these issues, they proposed that only selected features should be included for learning the classifier. Use of hybrid classifiers for classification as they inherit the properties of the base classifiers. Lastly select accurate classifier for better results. The proposed system is combination of different mechanisms [25].

In another paper Shengyi Jiang et. al. recommended an improved KNN algorithm for texe categorization. Author, in their research, combined KNN algorithm with one pass clustering algorithm to construct a model for text categorization. KNN algorithm itself is considered as slow learner and less effective classifier for text classification. But its combination with clustering algorithm produced improved results. Experiments conducted on three different datasets showed that text similarity is reduced considerably. Results compared with other classifiers like SVM and NB shows that this proposed model outperform advance classifiers like SVM and NB [26].

In another study Huy Nguyen Anh Pham and Evangelos Triantaphyllou proposed a meta-heuristic approach to improve the performance of classification algorithms, which they called as "Convexity Based Algorithm (CBA)". The proposed approach fist finds out the total misclassification cost. Then it uses this cost as a weighted function of penalty costs and corresponding error rates. Training data is then divided into different regions. Then classification tools are used in

combination with CBA. The experimental results show that this proposed system can fill gaps in current techniques [27].

There are numerous applications of text classification, for example, product reviews, mail directing, research paper categorization, email filtering [28], content arrangement, news categorization [29], spam filtering [30], automatic essay grading [31],sentiment analysis [32] and narrow casting. Keywords are used for this purpose. Keywords are extracted from original text to classify the document. Keywords contains most substantial words that contain most important information about text data. These keywords are then processed to classify a document.

TF-IDF and WordNet. TF-IDF algorithm are used to get keywords from documents. The words which have highest similarity are taken as keywords. KNN, NB and Decision Tree algorithms are use and their performance is compared. The performance of Decision Tree algorithm was better than others.

2.4.2 Ensemble classification

Hossainet. al. in their study; “Can ensemble techniques improve coral reef habitat classification accuracy using multispectral data?” Tested some classification techniques, they used common methods like; maximum likelihood, K-nearest neighbor, minimum distance, they compared these methods with ensemble classifiers. They showed that accuracy per-class of the surroundings discovery enhanced in the ensembles, specifically, Majority Voting classifier attained95%, accuracy for coral, 65%, for sparse coral, 75% for coral rubble and 95% accuracy for sand. Ensembles enhanced the accuracy by 28%, as compared to traditional methods. Therefore, the ensemble methods can be favored over the conventional techniques [33].

Some other scientists driven for the development of order calculations utilized “Genetic Algorithms (GAs)” (Goldberg, 1988) which join presence of fitting string structures using an organized however randomized data interchange to frame a hunt calculation. The subjected calculations have been utilized in characterization methods, AI and information mining applications [20] [21]. Hereditary Algorithms have additionally been utilized in enhancing some other learning strategies, as neural systems [22].

Joshua S. White et. al. proposed “Hybrid Sentiment Analysis Utilizing Multiple Indicators to Determine Temporal Shifts of Opinion in OSNs”.An event on social media can be classified on the basis of accurate understanding of what subjects are related to the event. In many cases, the significant evidence of key players of event also play important role in classifying the event. there

are largely two methodologies used for sentiment analysis of qualitative text data. NLP and bag of words are usually used for this purpose. Results in both cases are very much similar. i.e. around 80%. Author, in presented this paper presented a hybrid approach of the two techniques mentioned above, which produced better results. [34].

Sharareh R. Niakan Kalhori1 and Xiao-Jun Zeng, proposed "Improvement the Accuracy of Six Applied Classification Algorithms through Integrated Supervised and Unsupervised Learning Approach". They offered a combined methodology on the basis of supervised and unsupervised techniques to increase the accurateness of six classification algorithms which are established to forecast consequence of tuberculosis treatment development but their accurateness requests to be improved as these methods are not accurate as much as required. The proposed method: "integrated supervised and unsupervised learning method (ISULM)" was applied on more than six thousand TB patients. They first applied six selected algorithms on this data. then has a new way to increase model accuracy. The corpora of six thousand four hundred and fifty Iranian TB patients in DOTS treatment they applied Integrated established schemes with k-mean clustering algorithm to compute result of TB treatment. The proposed system showed increase in accuracy for all applied classifiers ranging between 4% and 10% [35].

2.4.2.1 Sentiment Classification using homogenous ensemble:

Ensemble classification is considered a better approach and is widely used in classification. Ensemble classification refers to combine different classifiers predictions to classify a new sample. There are two types of ensembles. i.e. Homogenous classifiers and heterogenous classifiers. *Homogeneous* ensemble uses same type of classifiers at base level. In this chapter, we use homogenous classification methods to classify our data sets. Bagging and boosting are two popular methods used for homogenous ensemble classification.

Many researchers have been interested in ensembles learning techniques, Verma and Mehta in their research work, have suggested a new ensemble learning method Bagging and Boosting for the classification of the five UCI datasets from the field of Bioinformatics with a suitable base classifier. They conducted experiments using Weka and Java Eclipse and showed that their method gives better precision with less root mean square error rate using the technique of ensemble learning. They concluded that their proposed ensemble learning technique is more appropriate in

providing the solution of the problem of classification in bioinformatics field. Such techniques can be used efficiently in related situations of classification dominion [36].

Fernando Enríquezet. al. in their study; "A comparative study of classifier combination applied to NLP tasks" presented a comparison of classifier grouping techniques, which have been positively practiced to several tasks which includes Natural Language Processing. As per their research, range of classifier grouping methods but the most important trouble is to select the one which is at top and appropriate for a specific task. In their study they evaluated the various techniques of grouping and their performance for example voting, Bayesian merging, behavior knowledge and bagging, for the part-of-speech classification using nine datasets in five different languages. They showed that some techniques which are not very prevalent, but these techniques could establish considerably enhanced performance. Furthermore, they showed how the size of a dataset and quality effect the performance of combination techniques. They also provided the results after applying the combination techniques to the Natural Language Processing tasks [37].

R. Mousavia et. al. projected "improved Static Ensemble Selection (SES) using NSGA-II multi-objective genetic algorithm called: SES-NSGAII". First method in the early stage nominate the finest classification algorithms and their combiner, by instant inaccuracy improvement and variety targets. At next stage, the method of "Dynamic Ensemble Selection-Performance (DES-P)" is ranked up by means of the optional method of the first stage. The other obtainable technique in the subjected research is a mixture of technique which uses the capabilities of both the scientific method and is known as "Improved DES-P (IDES-P)". So, integration inert and dynamic ensemble techniques with using NSGA-II. Results of their study shows that the methods planned overtake the other ensemble techniques in relations of categorization correctness over fourteen datasets [38].

Fatemeh Nemati et. al. has built up a three-phase half breed information mining method of highlight determination and outfit learning grouping calculations. In their system at first level, the information gathering, and pre-handling is finished. At the second level, four "Feature Selection (FS)" calculations are utilized that incorporate chief part investigation (PCA), hereditary calculation (GA), data gain proportion, and help quality assessment work. Requirements setting of these strategies relies on the exactness came about because of the implementation of the "support vector machine (SVM)" calculation. At that point in the wake of choosing the reasonable model for each component, they are connected to base and outfit calculations. At this stage, the best

Feature Selection (FS) calculation is determined with its requirements setting is determined for next stage which is; application of the proposed model. At third stage, the calculations are connected for the dataset composed from every FS calculation. The experimental results verified that at the second stage, the best feature selection (FS) calculation is PCA. At the third stage, the artificial neural system (ANN) ada boosting (AdaBoost) technique has more accuracy [39].

Pooja Shrivastava and Manoj Shukla, performed comparative bagging analysis, comparative stacking analysis and comparative analysis of random subspace algorithms. In their study they compared results of forest fire data set using the subjected algorithms into WEKA suite. On the bases of the experimental they show by comparing these three techniques that there is improvement in accuracy of classification. The results of experiments show that combination of classifiers performs far better than the results produced by the other methods of classification. At the end, they explained the technique used to combine the classifiers to increase accuracy on the minimum time data set and reduced error rate [40].

M. Eftekhari & R. Mousavi et al, proposed a novel ensemble learning technique upon the basis of hybridization of ensemble classifier selection methodologies. However, system of ensemble learning is a method which increases the enactment of the classification algorithms but one of the important and basic challenges in ensemble learning is to combine the yields of base classifiers. M. Eftekhari & R. Mousavi in their paper, proposed an optimized approach named as “Static Ensemble Selection (SES)” and this is based on the NSGA-II multi-objective genetic algorithm. This algorithm is also known as SES-NSGAII. This algorithm chooses the finest classifiers alongside their combiner, through real-time reduction of fault & diversity objectives. At other phase, the “Dynamic Ensemble Selection-Performance (DES-P)” is enriched with consuming the firstly recommended technique. The recommended technique is a hybrid method that uses the capabilities of both DES & SES methods and it is named as “Improved DES-P (IDES-P)”. The main contribution of this research is to combine dynamic & static ensemble techniques and making use of NSGA-II. The result of their study confirms that the suggested technique performs best in terms of accuracy of classification as well as it is far better than the other approaches when tested over 14 datasets [41].

Thabit Sabbah et. al. anticipated “Hybridized term-weighting method for Dark Web classification”; the statistical approaches’ performance is restricted because of the insufficient accurateness formed by the incompetence of these approaches to comprehend the texts produced

by humans. Thabit Sabbah et. al., in their study, presented a hybrid technique for feature selection depends on elementary term-weighting methods for precise discovery in textual perspectives of terrorism activities. The proposed technique combines the sets of features which are chosen on the basis of various individual feature selection approaches in to one feature space for better webpages cataloging. The performance of proposed technique is verified on a chosen set of data taken from a Dark Web Forum using different renowned text classifiers. They showed in the course of experimental findings that the proposed method has identified the terrorist activities more efficiently and left behind the single techniques and comparatively feasible in the number of features used for classification with privileged levels of hybridizations. Moreover, the hybridizing function brings into surface that feature set's dimensionality is significantly decreased by enforcing the Symmetric Difference function to combine the feature sets [42].

Gang Wang et. al. worked on Sentiment classification using ensemble learning. With the prompt increase and advancements in information technologies, contents generated by users can be appropriately sent online. Individuals, industries, and managers are involved in assessing the emotions behind the content, but there is no consensus on which emotion technique for classification is best. Current studies have shown that ensemble techniques can potentially be applied to sentiment classification. In this study, three common integration techniques "bagging, augmentation, and random subspaces" depend upon on five basic learners (decision tree, simple Bayesian, K nearest neighbor, maximum entropy, support vector machine). Performance was compared and evaluated. Used for sentiment classification. In addition, 10 datasets of national sentiment analysis were also analyzed to assess the efficiency of ensemble learning method in sentiment analysis. In this study which is based on 1200 comparative experiments performed on groups, observed results illustrate that the integrated technique can extensively progress the ability of emotion classification for a single basic learner. Among these three integration methods, the random subspace method shows improved results of comparison, but it is not so often described in the literature. Results of these experiments indicate that ensemble learning can be used as a practicable technique for classification of sentiments [43].

2.4.2.2 Sentiment Classification using heterogenous ensemble:

Another useful technique in supervised machine learning is to learn the combination and patterns of predictions of base level several classifiers. The purpose behind is to obtain higher level of

accuracy at meta level than individual classifier. Stacked generalization is the technique which is usually called stacking that deals with process of learning meta level classifier and to combine results predicted by different base level classifiers. Achievement of stacked generalization is due to diverse classifiers used at base level.

Eun Sung Lee in his study investigated the performance of stacking technique for estimation of the elderly depression on dataset from "Korea National Health and Nutrition Examination Survey (KNHANES)" for the duration of five years i.e. 2010 to 2015. The KNHANES is a freely available dataset intended at evaluating the dietary and health standings of Korean people since 1998. It is a country wide illustrative survey comprising about 10,000 persons every year as a survey sample. Eun Sung Lee in his study examined the modifications in the performance of stacking technique when merging 5 (five) classifiers that is, "NN, DT, LR, NBN and SVM", the enactment of stacking technique is measured in terms of accuracy and area under the curve which displays vigorous pattern when the classifier especially the base level classifier is comparatively trouble-free for example DT or LR, and the meta level classifier is somewhat multifaceted for example "SVM, NN or NBN". [44]

Many combination techniques are suggested, but till now it is challenging to find the appropriate combination pattern. Model Stacking is a technique having several approaches to group the predictions of classifiers learned to produce a new predictor for predictions as new instances occur. Researchers have proved that model stacking is generally more precise as compared to any other classifier individually used. Anjali Gupta and Amit R. Thakkar proposed that meta-heuristic techniques can be considered as a way out to discover improved formations. "Genetic algorithms" and "Ant Colony algorithms" are famous methodologies upon which present studies are around. Anjali Gupta and Amit R. Thakkar worked on previously used meta-heuristic methodologies for the optimization of stacking formation and they tried to figure out the future work that can be done to incredulous the limitations of existing methods. [45].

As compared to models stacking, ensemble method is a process to make a classifier by applying different algorithms to a single dataset. Complex techniques are typically used in this setting for the combination of classifiers. Stacking is a quite simpler model which is often used to combine methods [46] [47].

Ensemble of categorization algorithms, or a valid assemblage of dissimilar classifiers, regularly generates better categorizations as contrast to featured classification algorithms. Although, the

problem about which classification algorithm should be chosen for an exact situation to generate a perfect assembly has been examined again and again. In addition, ensemble method is often experimentally more costly as it needs the execution of several morpheme for a single job. To offer answer of such troubles, Dan Zhu et. al. projected “a hybrid technique to choose and merge the models to construct ensemble classifiers by incorporating stacked generalization and Data Envelopment Analysis” [48]

Georgios Paliouras et. al. explained ability of determination and stacked generalization. Another structure is prescribed which sets up well known strategies for data extraction with stacking. To create an informational collection at meta-level that contains highlight vectors they did cross-approval on the information index which is basically base level informational index that comprises content reports set apart with related classifications or classes. At that point a classifier is found out utilizing the new vectors. Henceforth, base-level data extraction strategies are joined with a typical classification algorithm at the meta level. Discoveries of explained investigation show that both casting a ballot and stacked generalization are enhanced, wherease utilizing probabilistic gauges by the base-level strategies. Stacking, indicated continually compelling similarly or preferred in pretty much every area over casting a ballot and consistently enhanced than the finest base-level techniques [49].

Combined characterization strategies commonly construe all classes of a social informational index, by methods for the proposals almost any class mark to influence recommendations about associated class. Kou and Cohen presented a powerful method based on stacking procedure that has similar precision to additional refined and blend grouping approaches. While with trials on both the fake and genuine information, they demonstrated that the fundamental purpose behind the better execution of the stacked model is the decreasing in partiality from erudition the stacked model on proposed classes as opposed to real classes. Additionally, they uncovered that the consolidated characterization strategies and stacked model’s execution can be perceived to a suggested weighting of social highlights at preparing stage [50]

Riyaz Sikora et. al. proposed an "adjusted stacking troupe AI calculation utilizing hereditary calculations". They utilized number of data sets taken from UCI Data Source for their investigation. They utilized five learning calculations in the stacking calculation which are “OneR, IBk, NNet, Naïve Bayes, and J48”. The got finest improvement in execution on the Chess set, where the adjusted stacking calculation had the option to build the exactness of forecast by over

10% when contrasted with the typical stacking calculation. The preparation time of both the calculations was likewise considered; as a rule, the adjusted stacking calculation devours additional time as analyzed to normal stacking calculation as it incorporates the running of GA. They likewise recommended that for preparing time reduction can be accomplished by parallel learning of individual learning calculations [21].

Paul N. et. al. in their study; "The Combination of Text Classifiers Using Reliability Indicators" introduced a probabilistic model for combining the data mining classifiers which is based on the context-sensitive consistencies of classifiers to be used in the combination. Due to the perception that different classification algorithms act qualitatively different ways it has been a long-interested attempt to construct an enhanced meta classifier through some grouping of classifiers. The technique combines the reliability indicator variables that deliver signs about the performance of classifiers in various circumstances. Paul N. et. al. presented measures for the construction of meta classifiers which consider reliability signals in addition to classifier outputs, they also provided an analysis of comparative studies used to assess the technique [51]

Eitan Menahemet. al. proposed; a novel system which is constructed from three layers combination of classifiers. Stacking is a technique in which different base classifiers are joined to create one meta-classifier which is learned on the outputs of base classifiers. Such a methodology offers certain benefits for example, plainness; performance which is comparable to the best classifier; and the proficiency of joining classification algorithms encouraged by dissimilar inducers. The multiclass problem is one of the short coming of stacking, it appears to perform poorer than other approaches of meta-learning. Eitan Menahemet. al. in their research presented; Troika, which is a novel stacking technique for refining ensemble classification algorithms. They tested their novel technique on numerous datasets their results demonstrated the preeminence of the suggested technique to the previous ensemble systems such as; Stacking and Stacking C, particularly when the categorization task comprises of more than two categories [52].

Jin Chenet. al. worked at the hyper spectral data for the improvement of the classification accurateness using "support vector machines (SVMs)" with stacking technique and the corresponding material of size and form feature sets. Their proposed technique is as following: SVMs classifiers are learned and accepted in size and shape feature sets at base level (termed as base level SVMs); then results of base level (decision values) are forwarded as inputs of meta level classifiers. Since the results of base level SVMs carry more information than class labels so, meta

level classifier accepts Support vector machine's (meta level SVMs) learned in the meta level feature space. Furthermore, they also discussed the possibility of decreasing the number of base SVMs by meta-level feature selection and presented a modest solution. They did experimentations on a standard data set and established, their technique considerably did well as compared to the techniques with the solo feature sets and other merging techniques, i.e., voting, unconditional maximum decision value, and stacking with labels of class [53].

Shoushan Liet. al. presented a classifier grouping methodology for sentiment analysis. Firstly, various classifiers are produced through training the data with number of features like unigram and some POS. Then, for the next step classifier selection method is used to select a part of the classifiers for the combination. At Final step, the selected classifiers are joined using numerous combining rules. Their experiments show that all the combined arrangement or methodologies with different rules of combinations outperform single classification algorithms and the sum rule achieved the best performance with an enhancement of 2.56% over the best single classifier [54].

Ireneusz Czarnowski and Jędrzejowicz proposed stacking technique for the improvement of procedure of combined mechanism for classification and guarantee variation between prototypes. Combined mechanism for learning is assumed as an incorporation lessening of data through the learning procedure. Combination like this allows to present alteration mechanisms inside learning procedure by amendment of the data with to find the better illustration from the in the context of the learning performance standard. Data amendment can be done through data reduction in both scopes, that is; the features and instances generating the set of models. At present, the reduction of data has become an important methodology for doing analysis of big data and the results enhancement of the mechanism of learning. To authenticate the proposed approach in this study authors have carried-out computational experiments. Their study contains the explanation of the methodology and the conversation on the validation of experimental results [55]. The focus of Ireneusz Czarnowski and Jędrzejowicz research was on using rotation-based technique and stacking by decreasing data to improve performance and generalization ability of classifier. The objective was to gain required necessary information from a high dimensional data to train a classifier, particularly when the data rates are high. Their study explains that applying machine learning combiner with merger of stacking and rotation-based ensemble methods based on decreased data improves the accuracy classification process. Their experiments confirmed their findings. [56].

Ireneusz Czarnowski and Piotr Jędrzejowicz, in another study proposed that reduction of data can upsurge generalization capabilities of training model and reduce learning time. It can also be helpful especially in examining big data sets. In their study authors are attentive on the machine learning e.g. with the reduction of data. In their study the reduction of data is done by selecting related instances, called models. A modeled methodology based on the hypothesis that model selection is performed by a team of agents. Each instance (prototype) is selected from a cluster of instances. Every single prototype is obtained from each cluster. They used a kernel-based fuzzy clustering algorithm for cluster initialization. Combination of data reduction and a stacking is key feature of this study. To validate the proposed methodology, they performed computational experiments. They also experimentally evaluated the results of clustering methods and the number of stacking folds used on classification accuracy [57].

Yuan Tian et. al. worked on classifier combination for wheat leaf diseases recognition. Proper detection of disease using pattern recognition and computer vision has been considered and examined to decrease the loss, as well as in conclusion attain healthy farming. Yuan Tian et. al., in their study proposed a novel approach of Multiple-Classifier Scheme based on "support vector machine" for the purpose of detection of disease in wheat leaves categorization for greater accuracy of detection. The dataset of images used in the study were taken before a black uniform background of infected leaf samples with Puccinia Striiformis, Corrode Puccinia Triticina, Leaf Blight, Powdery Fungus. Three feature sets comprising of color, shape and texture were created for the classification. The suggested arrangement scheme was based on stacking and comprised of two stage structure: first one is base-stage and the second one is meta-stage, where base-stage was a component of three types of SVM-based classifiers trained by three sets of features and meta stage is one unit of SVM based decision classifier learned by meta-feature set which are produced by a novel data fusion mechanism. The approach proposed by Yuan Tian et. al. is further flexible as well as it has greater success rate of recognition as compared with other individual classifiers and other schemes of ensembles classifier for the detection of wheat leaf infections [58].

Hybrid algorithm of classification for prediction of terrorism is used in North Africa & Middle East. Machine learning techniques used for support in decision & prediction are very famous to talk about nowadays. Approaches for hidden learning, non-symbolic information delivers improved accuracy in predictions. In any case, learning unequivocal methodologies, representative knowledge gives in increasingly intelligible models. Hybrid machine learning models join

qualities of both knowledge portrayal of model kinds. In this examination creators looked at prescient precision and unambiguousness of express, certain, and hybrid machine learning calculations. This exploration depends on anticipating fear based oppressor bunches capable of attacks in North Africa and Middle East from year 2009 up to 2013 by looking at various norm, gathering, hybrid machine learning methods in particular; Naïve Bayes, K-nearest neighbors, Decision Tree, Support Vector Machine; Hybrid Hoeffding Tree, Functional Tree, Hybrid Naïve Bayes with Decision Table, Classification by means of Clustering; Random Forests; and Stacking classifiers. A short time later look at the outcomes acquired from leading the analyses as per four distinctive execution measures. Investigations were completed utilizing certifiable information portrayed by "Global terrorism Database (GTD)" from National Consortium for the "investigation of terrorism and Responses of Terrorism (START)"[59].

Christiane Lemke et. al. worked on Meta learning which appealed significant attention in the machine learning research community in the last many years. However, some difference remains on what does or what does not establish a Meta learning problem, and in which backgrounds the word is used in. This review targets at giving an all-inclusive summary of the research guidelines followed under the heading of Meta learning, reconciling various definitions given in scientific literature, cataloging the ranges involved when scheming a Meta learning structure and finding some of the upcoming research challenges in the domain of Meta learning [60].

The collective classification technique uses inference from a category tag to influence inference from related category tags, so it collectively infers all category tags from a relational dataset. Kou and Cohen presented an efficient stack-based relationship technique that, while simple, has the same precision as more complex joint reasoning methods. Through experiments with real and artificial data, they show that the core reason of the degradation of the performance of the stack model is to reduce the bias caused by learning the stack model in inferred labels instead of real labels. The decrease in variance because of conditional inference also helps the effect, but the effect is not as strong. Furthermore, they show that the joint reasoning and performance of stacked algorithms can be attributed to the implicit weighting of local and relational characteristics during learning. [61].

Relative investigations of text classifiers have been done by different scientists. An investigation of customary content classifiers shows that a few classifiers are acceptable text classifier in a binary setting. Be that as it may, record grouping regularly includes numerous classes or

classifications. This prompted the improvement of the multi-class calculations, yet intricacy was relative to the check of classes present. This caused the preparation time of calculations to rise quickly. A few calculations cause a high-test time multifaceted nature which before long gets unfeasible with huge sets. In the current day, order speed is a significant part of report arrangement alongside characterization exactness. Be that as it may, there is no reasonable champ on the two includes in a multiclass order arrangement. In the current time, the web has brought about an immense measure of information alongside scientific categorizations comprising of thousands of classifications. Conventional single classifiers are presently incapable to deal with the degree and complexity of this information with required degree of accuracy.

The causes of classifier mix research appear to be established in design acknowledgment where the difficult area comprises of high measurements however modest number of classifications. As such the vast majority of the techniques talked about are gathering strategies where every classifier is applied on the full information space. As talked about before, this isn't appropriate for the current content space with an enormous number of classes at single and numerous levels.

The utilization of feature subsets has been proposed in a writing to lessen classifier computational complexity and training time. Be that as it may, there are no unmistakable pointers regarding the manner in which the original feature space can be parceled. Arbitrary projections of the feature space have been utilized yet there is no assurance that these projections contain essential distinctive qualities. A partition of the feature space which identifies with the parting of the hidden information is required logically. Tulyakov et al. have suggested that the ideal method is segment the feature space into the territories or categories related to various classifications [62].

The improved precision of ensemble algorithms, in light of model change decrease and less significantly inclination decrease, is built up on the humble yet compelling process of gathering averaging or broadly held vote. For instance, the relationship of the decision process of a specialist board can show the instinct behind ensemble strategies. A wiped out patient counsels a gathering of free clinical experts and the gathering decides the analysis by dominant part vote. Most eyewitnesses would concur that the patient got a superior or progressively precise analysis when contrasted with one got from a solitary pro. The precision of the finding is presumably higher and the fluctuation or error of conceivable misdiagnosis is lower, albeit greater understanding doesn't generally suggest exactness [62].

Data splitting includes partitioning the data into an explicit training dataset used to set up the model and a discreet test dataset used to assess the model's presentation on hidden data. It is valuable when you have an extremely huge dataset so the test dataset can give an important assessment of execution, or for when you are utilizing moderate strategies and need a brisk guess of execution. There are various strategies that you can use to assess the accuracy of your model on concealed data. For instance, Data Split, Bootstrap, k-fold Cross Validation, Repeated k-fold Cross Validation, and Leave One Out Cross Validation. Scientist utilize simple random sampling when the sampling frame is steady and can be analyzed at a fixed point in time, for example, clients' approaching email messages during the main week following the dispatch of another item. In simple random sampling, every occurrence has an equivalent probability of incorporation in the example and that probability depends on population size. For instance, if the researcher distinguished 963 tweets containing a watchword sent on a given day, simple random sampling of those tweets would permit each tweet a 1 out of 963 possibility of determination into the example. Random sampling with and without replacement can be led inside an enormous informational collection by means of standard statistical analysis.

A stratified random sample is acquired by isolating the population components into groups, or layers, with the end goal that every component has a place with one and only one stratum, and afterward freely choosing a straightforward random sample from every stratum. This sample review configuration has three significant favourable circumstances over basic random testing. To start with, the variance of the assessor of the population mean is typically decreased on the grounds that the variance of observations inside every stratum is normally littler than the general population variance. Second, the expense of gathering and examining the information is regularly decreased by the detachment of an enormous population into littler layers. third, separate appraisals can be gotten for singular layers without choosing another sample and, subsequently, without extra expense. A stratified random sample is one gotten by isolating the population components into nonoverlapping gatherings, called layers, and afterward choosing a basic random sample from every stratum. Many researchers use Stratified random sampling to overcome bias in splitting of data. Stratified random sample is used in data splitting function of R in this research. [63] [64].

In customary ensemble learning, we have various classifiers attempting to fit to a training set to estimate the objective capacity. Since every classifier will have its own yield, we should discover a consolidating system to join the outcomes. This should be possible by casting a ballot (greater part wins) or weighted deciding in favor of (model a few algorithms have more weight than the

others), averaging out the outcomes, and so on. This is the conventional method of ensemble learning.

Another assessment proposed a hybrid arrangement method for Arabic text mining that solidifies the advantages of statistical classifier (NB) and rule-based classifier (AC) in one framework and endeavored to vanquish their obstructions. In the essential stage, the important order rules are found using AC approach and confirmation that related features are fundamentally identified with their classes. In the ensuing stage, the NB is utilized at the back completion of disclosure system and takes the discovered rules, interfaces the related features for every characterization and masterminds' texts based on the statistical data of related features. The proposed method was evaluated on three Arabic text datasets with various classes with and without methods of feature selection. The investigational results showed that hybrid method beats AC independently with/without feature selection methods and it is better than NB in very few cases just with some feature selection methods extraordinarily when chosen feature subset is small.

Whereas in stacking, the mechanism of combination is that, results of the classification algorithm (Level 0 classifiers) will be used for the training of another classification algorithm (Level 1 classifier) to estimate the same target function. Basically, you let the Level 1 classifier to figure out the combining mechanism this is what we are focused on. We focused on model stacking for our tests. In this method we basically build several models of different types and then chain the predictions of these principal models to build a novel supervised model which acquires how finest combination of those predictions can be made.

The issue of programmed grouping of scientific texts is thought of. Strategies dependent on statistical analysis of probabilistic distributions of scientific terms in texts are talked about. The methodology for choosing the most instructive terms and the strategy for utilizing assistant data identified with the term's positions are introduced. The aftereffects of trial assessment of proposed calculations and techniques over genuine world Scientometrics data are accounted for.

2.5 Discussion:

Number of studies has been conducted to improve the classification of text data. quite a few includes reviews on different products and services. Few of these has been discussed below.

Xu, Kaiquan, in his research proposed a novel graphical model to visualize relative relations between products from customer reviews. He tried to discover the interdependencies among products by evaluating customer reviews. His main objective to discover potential risks and to help

enterprises to further design new products and help them making new marketing strategies. He experimented on a corpus of Amazon customer reviews and his model shows more accurate results than the standard approaches [63].

The growth of e-commerce has resulted in enormous growth of customer reviews on e-commerce websites and apps. A potential customer usually goes through a lot of customer reviews before making up his mind. However, mining an information of customer's interest is still an uphill task. Miao, in his study proposed a text mining system which draws out useful information from customer product reviews. A novel ranking system taking temporal opinion quality (TOQ) and significance into account is created to meet client's data need. Exploratory outcomes on a certifiable informational collection show the framework is plausible and viable [64].

Elli, Maria used MNB and SVM classifiers to classify Amazon Reviews. They also use machine learning techniques for fake review detection and trend patterns. The aim of this project is to extract sentiment from more than 2.7 million reviews and analyze the implications they have in the business area. The final results show MNB 72.95% and SVM 80.11% accuracy. Classifiers performance for both cases is very high [65].In this section, a comparative analysis of other related work with the proposed model is presented. This comparative analysis is based on accuracy measure. The table below shows the comparative results on similar type of data. [66] [67] [68].

Table 2-1 Comparison of previous literatures

Paper Title	**Dataset**	**Classifiers**	**Accuracy %**
Elli, Maria Soledad, and Yi-Fan Wang. "Amazon Reviews, business analytics with sentiment analysis." (2016):	Review of cell phone & accessories	MNB SVM	73 80
Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches Using Ensemble Learning & Voted Algorithm Publisher: IEEE Rubeena Parveen ; Neelesh Shrivastava ; Pradeep Tripathi Published in: 2nd International Conference on Data, Engineering and Applications (IDEA) 2020	Movie Reviews	NB SVM RF	86 84 88

A Multiobjective Weighted Voting Ensemble Classifier Based on Differential Evolution Algorithm for Text Sentiment Classification Aytug Onan, Serdar Koruko ˘ glu , Hasan Bulut Expert Systems with Applications 2016	Camera reviews Music Reviews TV Reviews Laptop Drugs	Stacking with base classifiers LDA, NB, SVM, LR, BLR	83 83 87 95 82
Shaikh, Tahura, and Deepa Deshpande. "Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews.", (2016):	Review on books Review on music Review on Camera		70 70 80 62 80 68 62 80 68
Improving text summarization of online hotel reviews with review helpfulness and sentiment Chih-Fong Tsaia Tourism Management 80 (2020) 104122	Online Hotel Reviews	DT RF LR SVM	71 75 69 63
Enhancing the performance measure of sentiment analysis through deep learning approach Rajit Nair1, Vaibhav Jain2 and Preeti Sharma Nair3 (2020)	Movie Reviews Hotel Reviews	SVM MNB KNN RNN-LSTM SVM MNB KNN RNN-LSTM	90 87 85 93 87 84 82 89
Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques Apeksha Arun Wadhe. Shraddha S. Suratkar. 2020 International Conference on Industry 4.0 Technology (I4Tech) Vishwakarma Institute of	Tourist Place Reviews	CountVectorization + NB TFIDFVectorization + NB CountVectorization +SVM TFIDFVectorization +SVM	83 84 80 85

Technology, Pune, India. Feb 13-15, 2020		CountVectorization +RF TFIDFVectorization +RF	85 86
Proposed Model	IMDB Reviews Amazon Reviews YELP Reviews	Heterogenous Ensemble Method	84 80 93

Chapter 3 : Datasets & Pre-Processing

This chapter discusses the data set characteristics and the preprocessing techniques which are proposed for the improvement of classification algorithms as per one of the objectives of this study. First part of this chapter discusses the datasets, used in this study. When there is a discussion regarding data, one commonly thinks of particularly large dataset with a large number of instances and attributes, but this is not must case; there could be so many different forms of data e.g. Unstructured/structured tabular form, images, audio files, and videos etc. Dataset could be considered as a group of instances, these groups are frequently called as records, patterns, events, samples, observations, or objects. Instances or these objects are defined by several properties, which consist of basic features of an instance, such as the mass of a physical object or the time at which an event occurred, etc. Features are often called as member variables, properties, characteristics, attributes, or dimensions.

Second part of this chapter discusses the preprocessing steps used in this study. As machines are unable to understand or process the data in unstructured form. In process of machine learning preprocessing of data is the method through which we can transform it or to pass it to a state due to which machine can effortlessly understand it or in other words then data can be analyzed or interpreted by the algorithms. Also, by finding and removing noise words in advance, we can increase greatly the accuracy of our algorithms.

3.1 Internet Movie Database (IMDb)

Dataset of movie reviews from the Internet Movie Database (IMDb) are used in this thesis which is available on internet publicly. This data was accumulated by Andrew Maas et al. this dataset consists of Fifty Thousand reviews, Twenty-Five Thousand for training purpose and Twenty-Five Thousand for testing. It is quite a balanced dataset with equal number of training and testing samples. But we have converted this dataset as per our requirements. i.e. different samples of 10,000, 6,000, 3,000, and 1500 samples has been taken. For this study, we have taken samples of at least 300 characters and 70% data objects are used for the purpose of training while other 30% is used for the testing purpose. We selected these reviews as one of our datasets as it contains a widespread variety of social sentiments and covers most of the properties which are relevant to

sentiment classification. Since in most recent research, movie review data is selected for sentiment classification by a lot of researchers.

Subjected movie review dataset is a collection of reviews of movies re-claimed from the website; imdb.com in the early 2000s by Lillian Lee and Bo Pang. They collected these movie reviews then used it for their research on processing of natural language. Also, they made it available on internet for public usage.

The subjected dataset was initially used at large in 2002, then an update was released in 2004, which is referred as “v2.0”. This dataset consists of both negative and positive movie reviews taken from an archive of a newsgroup presented at IMDB. Hyperlink of this dataset can be found in Appendix A.

Figures 3.1 and 3.2 shows snapshots of sample IMDb dataset of negative and positive reviews respectively

Class	Review
0	"This is possibly the worst movie i've ever seen, it was horribly done it didn't flow it was v
0	"Why, oh why, is this trash considered a classic? I've seen higher body counts on episodes of '
0	"Following up the 1970s classic horror film Carrie with this offering, is like Ford following
0	"The plot sounds vaguely interesting ... a scientist (Bateman) discovers how to transplant ani
0	"And I repeat, please do not see this movie! This is more than a review. This is a warning. Th
0	"For anyone who has seen and fallen in love with the stage musical A CHORUS LINE, the movie is
0	"The Other Boleyn Girl - not to be confused with the book it claims to be based upon. This mov
0	"It's 2005, my friends...a time of amazing special effects and an age of technology. So, why c
0	"This movie came very close to being a good flick. The direction needs to be a bit smoother to
0	"Othello is set to burn the eyes of the viewers of this film. The bad depiction of Shakespeare
0	"This sequel to ""In the Heat of the Night"" will suffer in inevitable comparisons to its infi
0	"This may actually be the worst movie that I have ever scene. Incoherent would be a compliment
0	"I've just been at the cinema in down town Prague watching this film. Not due to t
0	"I was very excited when this series premiered in 2005. The premise was very simple and appeal
0	"This movie was in a sci-fi 50-pack a friend of mine got me for Christmas. It is very similar
0	"Horrendously acted and completely laughable haunted-house horror flick that has an out of pla
0	"I have to say this, this is the first movie I'm reviewing on here I didn't finish watching, I

Figure 3.1Negative Reviews IMDb sample dataset

```
1   "You don't have to be a Notre Dame football fan to enjoy this, because I am not....but, as a f
1   "This movie is well done on so many levels that I am in awe that the score is as low as it is
1   "I watched this film a few times in the 90's and nearly split my sides laughing each time. I l
1   "The Farrelly brothers, Bobby and Peter, are at it again. With ""Fever Pitch"" the creators of
1   "This multi-leveled thriller kept my attention throughout. It is disturbing and informative to
1   "This movie was very good. If you are one who likes to watch horror movies, I recommend it. Th
1   "I just saw this movie on HBO, and it was really good...a tragic love story indeed! I really a
1   "A sweet and totally charming film, Shall We Dansu? made me laugh and cry. At first appearance
1   "This really is a great movie. I don't think it has ever gotten the recognition that it deserv
1   "A long-defunct prison, shut down for over 20 years, is re-opened and Ethan Sharpe (the late,
1   "I think that most of the folks who have posted comments on this movie don't understand how to
1   "These slasher pics are past their sell by date, but this one is good fun.<br /><br />The vale
1   "I don't cry easily over movies, but I have to admit, this one brought me to tears. Although I
1   "I feel like I have some uber-rare disease that no one has heard of and I have finally come ac
1   """Unconditional Love"" starts with great promise. As directed by P. J. Hogan, the film works
1   "The charm of Otto Preminger's grandiose, visionary film noir is that it has ambiguous intenti
1   "*SPOILER ALERT: I wish I could discuss this without revealing specific plot points, but I can
1   "This is the episode that probably most closely relates to it's partner law, ""Thou Shalt Not
1   "The Color Purple is a masterpiece. It displays the amazing acting abilities of Whoopi Goldber
1   """Idiocracy"" is the latest film to come from Mike ""Office Space"" Judge, and it certainly f
1   "This movie is a very enjoyable homage to the Bogart and other detective films of old. Robert
```

Figure 3.2 Positive Reviews IMDb sample dataset

3.2 Amazon Product Reviews

Reviews on Amazon product is another dataset, used in this study. This data is very important for company selling the product, as it can have quite negative or positive impact on sale of products of company. Since these reviews are unstructured and mostly rated as stars from 1 to 5. i.e. 1 star for unsatisfactory or negative review and 5 stars for excellent or surly a positive review. So, there is a need to convert this data to two classes i.e. positive and negative. So, we considered 1 star as negative review and 5 stars as positive review. Also, we confirmed this arrangement manually randomly and we found it correct. We selected maximum of 10,000 sample reviews from this bulk of data for our experiment in our study. Hyperlink of this dataset can be found in Appendix A.

This data can be used in many areas of research related to customer behavior and their interest in buying product online. We converted this dataset according to our requirement. We constructed only two field data i.e. Reviews and class from a multi field data.

The dataset covers the reviews of customer's text together with metadata, which consist of three major constituents:

- A group of reviews that are written in the Amazon.com market and related metadata between 1995 and 2015. The intentions behind this dataset is to facilitate future study into the possessions and the development of customer analyses possibly together with how people express and evaluate their experiences of products at some scale.

- A group of reviews in multiple languages about products from number of Amazon markets, with the intentions to enable the customers' analysis perception of the identical products and broader customer preferences through countries and languages.
- A group of reviews which are non-compliant with Amazon policies. The intentions behind it is to offer a dataset for the purpose of research on distinguishing promotional or unfair reviews.

Figures 3.3 and 3.4 shows snapshots of sample Amazon Reviews dataset of negative and positive reviews respectively

Class	Review
0	"There are cheaper more professional paint sprayer guns at home depot. This thing is crap
0	"They probably are fine, but they cost to much. Have sent back one batch and am sending back t
0	i bought two of these lights back in February the first one I opened had only 7 LEDs working I
0	"I've had a whole house filter for years and the heavy carbon filters I normally get last 6 mo
0	"Do not waste your money. The filter worked great the first couple of weeks but then started
0	"The light itself works well. I'm using it in a basement mounted to a wall in the stairwell.
0	"I really wanted to like this lock, I really did. The pro's were: Easy install
0	Bought 2 of these after seeing commercial both hoses only lasted a month one blew up and the o
0	"I wish I could return the item. the steam is luke warm at best, nor does it hold as much wat
0	This is the worst packing and product I have received from Amazon...maybe ever! The candles we
0	The quality of this switchplate is very poor. It is a junky plastic cover and it wasn't primed
0	"I bought the last two in Amazon's inventory to fix my torchiere lamps. In both cases, the re
0	"Item is not an OEM replacement, but is very close to the original. Unfortunately, it's not c
0	Beware and read a little more. The picture is of a red quart can....the gloss marine spar var
0	This product is not the size as shown in the picture it is actually only like 2 feet tall...wh
0	"this thing was working as advertised, and then one weekend, about a little over a month from
0	"This is my second n120. The first one had a dial thermostat and worked well for 3 years and
0	"I tried this and the Oxygenics Evolution since they are the only heads on the market with the
0	"I was planning to buy this for my girlfriend who likes to read on bed but it disappoints me t
0	"Putting USB sockets on an electrical outlet is a great idea but Newer Technologies, a company

Figure 3.3Negative Reviews Amazon sample dataset

1	"I agree with most everything that has been posted about this product--i read about 500 re
1	"My model GSH25JFRFBB GE fridge recently started displaying the typical signs if board fai
1	"Ordered the Baldwin 0014.003 mail slot 'used' from Amazon warehouse, but it looks brand-n
1	I have an older floor Ott light and it requires the plug in spiral bulb and not the Edison
1	"First of all, I am an electrical engineer. I design computer networks. I am pretty pick
1	"This one is so much more sturdier than the previous ones we have purchased. My wife has
1	I wish I had found these sooner. It would have saved my wood floors from so many scratches
1	"I buy Northern ultra plush 3-ply jumbo rolls (4.5\\""wide x 5\\""dia) of toilet paper fro
1	"See my review on the similar 36\\"" model. I also have a 24\\"", and both of those sizes
1	"We recently renovated our bathroom and had purchased this style of Moen fixtures for it.
1	"I bought this countdown timer to automatically turn off our gas fireplace after the selec
1	"I absolutely love this shower head. I removed the diverter valve because I like a stronge
1	"The fan was as-advertised - one piece metal housing (builder put in plastic) with large s
1	"It was hard finding a screwdriver to fit the screws to put the batteries in, but the ligh
1	"As far as I know, there is nothing even close to this on the market. I have had a few
1	"I kept receiving an F-1 error message. After checking the internet, found a site that ex
1	It provide exactly what I was looking for in the way of a small work light for my milling

Figure 3.4 Positive Reviews Amazon sample dataset

3.3 Yelp Dataset

The third dataset that we used in our research work is Yelp, which is becoming a platform on which large and small companies can be ranked publicly and evaluated in a fair competitive environment. Many companies argue that the fact that Yelp's main source of revenue is advertising sales leads to conflicts of interest, suggesting that companies can make their own efforts to display on more search results and competitors' pages. Yelp denied any wrongdoing and pointed out that the comment filtering algorithm applies to everyone in the same way. From its perspective, advertising is a way for websites to make money, while providing free services that everyone can use.

The Yelp review data set has hundreds of restaurants open every day, and provides choices from thousands of companies. Yelp.com is a website that offers to provide comments about restaurants and small businesses from the public. The site has thousands of visitors every day, thanks to its reliable restaurant ratings and reviews. For past few years, Yelp has hosted the Yelp Dataset Challenge once a year, encouraging students and other enthusiastic data scientists to use Yelp's huge data resources to promote the restaurant review space. The Yelp dataset can be used to analyze how the ratings of a particular restaurant are affected by variables such as the sentiment of the review, the length of the text, and the number of useful, interesting, and dazzling votes received by the review.

Overall, the yelp dataset contains 5,200,000 reviews, out of which 174,000 are businesses, 200,000 are pictures, 11 municipal areas, and 1,100,000 tips provided by 1,300,000 users. In business.csv, there are over 1.2 million characteristics for example parking, availability, ambience and hours. Furthermore, there are accumulated check-ins for all; 174,000 businesses.

Datasets are used with 3000, 6000 and 10000 records after having some preprocessing steps which are discussed in the later section of this chapter for the training and for building our model. Hyperlink of these datasets can be found in Appendix A.

Figures 3.5 and 3.6 shows snapshots of sample YELP dataset of negative and positive reviews respectively

```
"Class" "Review"
0   "My usual place doesn't deliver so I decided to try their food. After waiting one hour, I re
0   "I stayed at Circus Circus and saw there was a coupon for a free coffee. They serve breakfas
0   "I placed an online order last night. The owner called me this morning to ask for my credit
0   "DO NOT GO HERE TO GET YOUR HAIR CUT
0   "I wish I could give this place less than one star.  Was in on a Saturday,  they are technic
0   "I cannot believe the lack of customer service in this store. At first I went to find a sale
0   "We should have checked the Yelp reviews before we went to The Noodle Shop at Mandalay Bay.
0   "Worst service ever experienced at a Chinese restaurant. I paid our (my friends and I) meal
0   "A oublier! Salle de resto en haut qui semble encore en construction. Commande d'un apéritif
0   "I will never, never go back to this store! I went to use 2 coupons and when the cashier ask
0   "The workers are the only decent aspect of this office. Living in the service area for 2+ ye
0   "I was recently traveling the region with my sister, who is visually impaired and purchased
0   "- Slowest service ever. The way they made the sandwiches was extremely slow. Even worse was
0   "#meritageurbantavern we saw advertisements such as this about their grand opening, so me an
0   "I had a very bad experience in the cosmetics section of this shoppers. A chubby girl with a
0   "My friends begged me to come here and I wish they hadn't. When we walked in, there was tras
0   "We are vegetarians and have ordered Chole Bathura, Samosa Chaat and Pani Poori. Nothing was
0   "Over priced! Went for a beard trim he barley took anything off... I had a full Kurt Russell
0   "This is the worst steakhouse I've been to. I came for the ""fine dining"" aspect, not the l
```

Figure 3.5 Negative Reviews YELP sample dataset

```
1   "We came here for their AYCE sushi for lunch today. There will probably be a wait, but it's w
1   "ATTENTION PHONECIANS: If you live in AZ and don't like Pete's than you might be a looser. Be
1   "I decided to switch up my nail place since the old place I was going to was really understaf
1   "I first gave Tea at 94 4 stars and I've been meaning to update this for ages. I love Tea at
1   "I saw how many great reviews Eco-Tint had here on Yelp, so I priced them out and made an app
1   "i love this place! Its right by my work and I go over and grab some food on the regular. The
1   "Waren hier zum Groupon Pasta All You can eat und es war einfach super! Kann den Laden echt n
1   "Took my CCW class here, bought a gun here, and also done transfers through CTR. Always a goo
1   "I think this is my favorite restaurant in all of Cleveland! Every time I've been it's been d
1   "This was one of our favourite finds in Edinburgh. We actually found it in an advert in the b
1   "This place gave my friend and I (after asking if it was our first time eating there) EIGHT s
1   "My partner and I went in to spend roughly 100 bucks on a shrub for our back yard.. We quickl
1   "Soooooo soooo good. We've been here maybe 6 or 7 times and every food option I've tried so f
1   "I love this hotel.  My friends and I stayed at Vdara for my bachelorette party weekend.  We
1   "I love St. George and every time I am visiting, I try and attend the Divine Liturgy services
1   "Stopped in 4 pm on a Thursday before the evening rush. Very friendly employees!
1   "I've wanted to try this place out for a while but never did because it's a bit of a drive fr
1   "This place has been on my pizza bucket list for a long time and I'm so glad I got to FINALLY
1   "De-fricken-licious. I had never tried crawfish, but I'm a big seafood lover, so I was excite
1   "What an adorable little idea. I was afraid they would go out of business when BevMo started
1   "We are hooked!!! Came in for the first time today Unsure of what to really expect I can say.
```

Figure 3.6 Positive Reviews IMDb sample dataset

3.4 Pre-Processing Steps (Data Cleaning)

The following preprocessing steps are used to convert our datasets in ordered form. In this section we have discussed the base classifiers and learning methods. We have also discussed the methodology for our experimental design that is used in analysis.

The data has been cleaned up fairly, for example:

- The dataset consists of only English reviews.

- All the text has been transformed to lowercase.
- Only those reviews are selected having length of at-least 300 characters.

3.4.1 Importing the Corpus

For the analysis of data in R the initial step is to import or get data into the RStudio, which provides number of ways to get data read; one of these methods is to read data by using function; read.table, this function reads a data file in the form of table and then converts it to a data frame, with cases corresponding to lines and variables to fields in the file.

3.4.1.1 read.table

The subjected function is the main method of reading tabular data into **RStudio**.

3.4.1.2 Other Option for Reading Data

read.csv function and read.csv2 function are similar to read.table excluding for the defaults. These functions are used to read 'comma separated value' files ('.csv'). Likewise, read.delim function and read.delim2 function are used for reading delimited files, with default to the TAB character as delimiter.

3.4.2 Split into Tokens

Tokenization is a process of splitting a text document to the chain of words. In this process all punctuation characters as well as non–text characters are replaced by tabs and spaces. In order to use a text document in our experimental process, we must tokenize this document into meaningful stream of words. And, then this stream is used for more processing. The set of all the words such obtained by tokenizing all documents then forms a dictionary.

Following are few terms and variables that may be used frequently in different algorithms and classifiers. Let set of all the documents(reviews in present scenario) be denoted by D and $T = \{t1, \ldots\ldots, tm\}$ be the dictionary(collection of all the words in these documents), i.e. the set of all different terms occurring in D, then the absolute frequency of term $t \epsilon T$ in document $d \epsilon D$ is given by tf (d, t).

3.4.3 Filtering, Lemmatization and Stemming

It is very important to reduce the extent of dictionary, as more the size of dictionary, more the dimensions to deal with. In order to deal with curse of dimensionality, we use filtering methods and lemmatization or stemming methods.

Filtering refers to remove the words that has little meaning in the text or has no impact on overall effect of sentence. For instance, the prepositions, articles etc. which are used for grammatical purposes but do not have specific meaning. Also, the words that have very little meaning or very rarely used can also be removed to reduce the dimension of matrix. So, these kinds of words must be removed from dictionary which has no relevance in actual information, also (index) term methods can be used to reduce number of words in dictionary.

Lemmatization methods are usually not used as these are time consuming and also error prone. So, instead stemming methods are applied more frequently.

Stemming methods try to build the basic forms of words, i.e. it converts plural's form to singular form and removing other affixes. A stem is group of words with similar meaning. Every word is represented by a stem after this process.

3.4.4 Matrices

After importing data in R environment, corpus still required to be altered to the term matrix, which is the typical input for training and testing the models that depend on the bag-of-words supposition that is; the word order is not important.

The transformation to a document-term matrix or as matrix is also controlled by the tm package and is generally a simple function call; DocumentTermMatrix( ) or as.matrix( ) with the corpus as a parameter and a list of other attributes. These attributes determine the data available to the model and therefore sway how the model is trained or tested.

matrix generates a matrix from the given set of values.

as.matrix attempts to turn its argument into a matrix.

matrix (data = NA, nrow = 1, ncol = 1, byrow = FALSE, dimnames = NULL)

as.matrix(x, rownames.force = NA, ...)

3.4.5 Data Frame Source

This function creates a base from a data frame.

DataframeSource(x)

An instance of class DataframeSource extends the class Source representing a data frame deducing each row as a document.

Running the example gives a nice long list of raw tokens from the document.

Table 3-1 Vectorizing the word frequency of a document

Trained Text	*Positive Label*	*Word Vectors*	*This*	*Place*	*Is*	*Good*	*The*	*Bad*
This place is good.	*1*		*1*	*1*	*1*	*1*	*0*	*0*
The place is good.	*1*		*0*	*1*	*1*	*1*	*1*	*0*
This place is bad.	*0*		*1*	*1*	*1*	*0*	*0*	*1*
The place is bad.	*0*		*0*	*1*	*1*	*0*	*1*	*1*
p(label=1)	*0.5*	*p(Word|1)*	*0.5*	*1*	*1*	*1*	*0.5*	*0*
p(label=0)	*0.5*	*p(Word|0)*	*0.5*	*1*	*1*	*0*	*0.5*	*1*

3.4.6 tm_map (Transformations on Corpora)

Provides interface to apply functions of transformation to dataset.

tm_map(x, FUN, ..., useMeta = FALSE, lazy = FALSE)
S3 method for class 'VCorpus'
tm_map(x, FUN, ..., useMeta = FALSE, lazy = FALSE)

3.4.7 Stop words

This function returns different kinds of stop words with provision of numerous languages.

stopwords(kind = "en")

There are number of available stopword lists e.g. SMART, Catalan stopwords, A character vector comprising of demanded stopwords.

3.4.8 Character Translation and Casefolding

Translate/transform characters in character vectors, in particular from upper to lower case or vice versa.

tolower(x)
toupper(x)

tolower and toupper convert upper-case characters in a character vector to lower-case, or vice versa. Non-alphabetic characters are left unchanged. Returns a vector of characters having same length with the same attributes as x

3.4.9 Term-Document Matrix

These functions creates or coerces to a document-term matrix or a term-document matrix.

TermDocumentMatrix(x, control = list())
DocumentTermMatrix(x, control = list())

An instance of class TermDocumentMatrix or class DocumentTermMatrix which contains a sparse term-document matrix or document-term matrix. The property weighting comprises the weighting applied to the matrix.

3.4.10 Remove Sparse Terms

This function removes sparse terms from a document-term or term-document matrix.

removeSparseTerms(x, sparse)

Those terms having at least a sparse percentage of that is terms occurring zero times in a record are removed from the term-document matrix which in turn results a matrix of those terms only which have sparse factor less than sparse.

3.4.11 Data Splitting functions

Training and testing partitions are generated by using the function createDataPartition while another function createResample creates single or more than one bootstrap samples.

createDataPartition(y, times = 1, p = 0.5, list = TRUE, groups = min(5, length(y)))

Simple random sampling is used for bootstrap samples, s.

3.5 Data Mining Methods for Text:

Most of text data is unstructured. Main goal of using data mining methods is to structure the unstructured text data. A structured data is certainly more affective and useful for user. Book indexes and library catalogues are well known structured documents.

However, to structure the documents manually is a very time-consuming task. This is the main reason, why these documents are not updated regularly and not available in updated form readily. Also, frequently changing information sources like World Wide Web are required processes and methods which can update it regularly and readily.

Classification and clustering are mainly used to structure text documents based on some keywords or categories. We have used classification methods in this study to classify text data. In next section we have described briefly classification method used.

3.5.1 Classification

Main objective of classification is to assign a predefined label or class to a query sample. For example, we are to classify a document with a label say, "science", "math" etc. Whatever method is to be employed, in data mining classification, we must first go through training process. i.e. we need to divide the historically available data into training and test set. Let's start with training set D = ($d_1, d_2, \ldots, d_n$) of documents that are labeled with a class value $l \in L$ (e.g. Math, Computer, Physics). Next, we need to build a classification model

$$f : D \rightarrow l \qquad f(d) = L \qquad (1)$$

which can assign the correct label or class to a new query document d. A random fraction of above-mentioned historic data is then taken as test set. It is used to measure the performance of model obtained by training the set D. The sample used for testing is different from the samples which were used in training. The number of documents that is correctly classified in relation to the total number of documents is called accuracy and is a main performance measure that is used in this study.

Different accuracy measures can be used in analysis. We used accuracy measure for comparison and analysis of different data sets and classifiers. i.e.

$$\text{Accuracy} = \frac{TP+TN}{P+N} \qquad (2)$$

3.5.2 Information Extraction

One of major task is to find and extract information from a natural language text. We cannot use natural language text directly for any kind of analysis. We must go through large amount of text to extract use full information.

Since we know, what kind of information is required, therefore, the main text is transformed into use full information by converting it into tokens and assigning attributes to them. Consider the example in which information is extracted from a CV. Suppose some professional skills are provided in a CV.

Professional Skills: Data scientist, hands on expertise in machine learning, have contributed in projects on big data, Part of development team on Artificial intelligence. Implemented ensemble methods in R Language. Also done analysis using statistical techniques of results. Created different models using stacking and achieved higher accuracy.

Extracted Information: Data scientist, Artificial Intelligence, Ensemble methods, Statistical techniques, R language, machine learning, Big data,

A series of process is required to extract information from a given text. The process may include tokenization, sentence segmentation, part-of-speech assignment, and the identification of named entities, i.e. skill names, experience and names of organizations. At next level, parsing can be applied semantically and syntactically on sentences and phrases. Higher level phrases and sentences must be parsed, semantically interpreted and integrated. Finally, all the information gathered in the form of tokens or words is gathered in a database. And now this database is ready for further processing. i.e. we can now apply machine learning algorithms or classifiers for training etc.

3.5.3 Performance Evaluation

For the last few years, classification algorithms for text have been evaluated on several standard document collections. It is clear from previous experiences that the performance also depends on the document collection. Talking about the relative performance of classifiers boosted trees, SVMs, and k-nearest neighbors usually performed better, while Naıve Bayes and decision trees are less dependable.

By the use of classical example of spam detection, the conditional probability P(A|B) that a given text document, i.e. review (B), is spam, i.e., deceptive (A), is equal to the conditional probability P(B|A), scaled by the marginal probability of P(A) divided by P(B) (7).

$$p(Spam|Text\ Document) = \frac{\text{p(Text Document|Spam)p(Spam)}}{\text{p(Text Document)}} \quad (3)$$

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 3.7 Confusion matrix

Where, a trained or learned text classification algorithm attempts to properly categorize the occurrence (true positives) or nonappearance (true negatives) of the label or class variable.

Chapter 4 : Base Classifiers for Text Classification

Classification refers to accurately predict the class of each query sample in the data. The historical data set is divided into training and test sets. A classifier is trained on training data and a model is formed. Then this model is tested on test data set and accuracy percentage is found. If this model is adopted on the basis of accuracy achieved, then this model is applied on any data sample provided and required class is predicted.

In this study three data sets are used. i.e. Yelp reviews, Amazon reviews and IMDB reviews. It is two class data set with positive or negative reviews. Different samples of dataset instances of each data set are taken and their preconditions are kept same. So that there will be minimum variation in their results due to preconditions after training of classifiers in their prediction. This data is taken from internet from their respective sites. Since data is selected randomly by classifier so there is a chance of little variation in the accuracy of same classifier on different data sets. Also, vocabulary of text data is another factor which cannot be dealt with in preprocessing. Despite these two factors, we have tried to keep preconditions same for all datasets to get uniform results.

Firstly, different base classifiers from different family of classifies are used to train and test on their different data sets of same kind i.e., text datasets with two classes each.

4.1 Base Classifiers:

Seven different classification algorithms from different family of classifiers are used in this study. It is to find which algorithm gives us best result and how we can improve these results.

Datasets used are Amazon Reviews, Yelp Review, and IMDB Reviews. We have tried to keep the precondition same for all datasets as discussed in previous chapter. Different samples of datasets (10000, 6000, 3000, 1500) instances are taken for each dataset splitting into 70% for Training set and 30% test sets.

All the below mentioned classifies are trained on these training sets and then tested on test sets. Following table shows different variations of classifiers and the classifiers we used as base classifier.

Table 4-1 Selected Base Classifiers

Classifier Selected	***Classifier of same nature or of same family***
Linear Discriminant Analysis (LDA)	Principle Component Analysis (PCA), Quadratic Discriminant Analysis (QDA), Regularized Discriminant Analysis (RDA).
Classification Tree (CTREE)	Iterative Dichotomiser 3 (ID3), C4.5, Classification and Regression Trees (CART), Chi Squared Automation Interaction Detector (CHAID).
Naïve Bayse (NB)	Baysian Network Based Algorithms.
K-Nearest Neighbour (KNN)	1-Nearest neighbour (1-NN), Fuzzy Nearest Neighbour (FNN), Fuzzy Rough set Nearest Neighbour (FRNN).
Neural Networks (NNET)	Auto encoder, Probabilistic Neural Networks (PNN), Time Delay Neural Networks (TDNN), Convolutional Neural Networks (CNN), Deep Stacking Networks (DSN).
Recursive Partitioning and Regression Trees (RPART)	R-Implementation of CART (Classification and Regression Trees).
Support Vector Machine (SVM)	SVM-Linear, SVM-Polynomial, SVM-Homogenous, SVM-Hyperbolic.

The classifiers used are **LDA, NB, RPART, KNN, SVM, NNET, CTREE.**
Following is brief description of the selected classifiers.

4.1.1 Linear discriminant Analysis (LDA)

Linear discriminant Analysis (LDA), Normal discriminant analysis (NDA), or discriminant work investigation is a speculation of Fisher's linear discriminant; a technique utilized in insights, design acknowledgment and AI to describe and isolates two classes. The subsequent blend might be utilized as a direct classifier, or, even more ordinarily, for dimensionality decrease before later grouping.

LDA is immovably related to assessment of vacillation (ANOVA) and backslide examination, which in like manner try to communicate one ward variable as a straight blend of various features or estimations LDA is moreover solidly related to head fragment examination (PCA) and figure examination that both of them look for direct mixes of elements which best explain the data. LDA expressly endeavors to show the contrast between the classes of information. LDA works when the estimations made on free factors for every perception are ceaseless amounts. When managing unmitigated autonomous factors, the comparable method is discriminant correspondence examination.

4.1.2 Classification Tree (CTree)

CTree makes a tree-like model which calculates the estimation of an objective variable by taking in basic choice standards deduced from the information places of interest. Decision trees group the models by masterminding them down the tree from the root to some leaf node, with the leaf node giving the classification to the model. Each node in the tree goes about as an analysis for some trademark, and each edge falling from that center point identifies with one of the likely reactions to the investigation. This system is recursive in nature and is reiterated for each sub tree built up at the new nodes.

4.1.3 K-Nearest Neighbors (KNN)

For order and relapse a non-parametric technique the K-Nearest Neighbour (kNN) is utilized in example acknowledgment. In the two cases, the information contains of the k closest getting ready models in the segment space. The yield depends upon whether k-NN is used for arrangement or relapse:

Both for order and relapse, a gainful technique can be utilized to name weight to the obligations of the neighbors, with the objective that the closer neighbors contribute more to the run of the mill than the more far away ones. For instance, an ordinary weighting plan remembers for giving each neighbor a weight of 1/d, where d is the parcel to the neighbor.

The neighbors are taken from a huge number of articles for which the class (for k-NN gathering) or the thing property estimation (for k-NN backslide) is known. This can be thought of as the planning set for the estimation; in any case, no express preparing attempt is required.

4.1.4 Naive Bayes (NB)

Naive Bayes classifiers are a social occasion of clear "probabilistic classifiers" considering applying Bayes' hypothesis with strong (gullible) self-governance speculations among the features.

Naive Bayes is a reasonable strategy for making classifiers: models that relegate class names to give models, tended to, for example, vectors of highlight respects, where the class names are drawn from some confined set. There is decidedly not a solitary calculation for preparing such classifiers, yet a social event of figuring dependent on a typical norm: all gullible Bayes classifiers recognize that the estimation of a specific part is independent of the estimation of some other fragment, given the class variable. For instance, a characteristic thing might be an apple in the event that it is red, round, and around ten centimeters in breadth. A Naive Bayes classifier considers every one of these highlights to contribute wholeheartedly to the likelihood that this common thing is an apple, paying little character to any possible associations between the covering, roundness, and division transversely over highlights.

4.1.5 Recursive Partitioning and Regression Trees (RPart)

Decision Tree is a predictive modeling instrument that has applications spreading over different zones. All things considered; decision trees are created by methods for an algorithmic strategy that recognizes ways to deal with section an educational file subject to different conditions. It is one of the most comprehensively used and sensible methods for administered learning. Decision Trees are a non-parametric managed learning technique utilized for both gathering and backslide tasks. The objective is to make a model that predicts the estimation of an objective variable by taking in direct decision rules incited from the information features. The decision rules regularly if else decrees. The more significant the tree, the more befuddled the principles and fitter the model.

4.1.6 Neural Networks (NNet)

Artificial neural systems (ANN) are figuring structures awakened by the normal neural systems that set up creature minds. The neural framework itself interesting AI calculations and procedure composite data inputs. Such structures "learn" to perform tries by contemplating models, everything considered without being changed with any errand express principles

An ANN depends upon a group of related units or centre points called counterfeit neurons, which straightforwardly model the neurons in a characteristic cerebrum. Every alliance, similar to the neurotransmitters in a characteristic character, can transmit a sign start with one counterfeit neuron then onto the accompanying. A phony neuron that gets a sign can process it and after that sign extra fake neurons related with it.

In ANN application, the sign at a connection between counterfeit neurons are an authentic number, and the yield of each phony neuron is enrolled by some non-direct restriction of the aggregate of its data sources. The connection between counterfeit neurons are called 'edges'. Counterfeit neurons and edges consistently have a weight that alters as learning continues. The weight increments or diminishes the idea of the sign at a connection. Counterfeit neurons may have an edge with the genuine target that the sign is possibly sent if the hard and fast sign crosses that limit. Reliably, counterfeit neurons are assembled into layers. Various layers may perform various types of changes on their wellsprings of information. Sign travel from the fundamental layer (the data layer), to the last layer (the yield layer), potentially in the wake of crossing point the layers on different occasions

The fundamental objective of the ANN approach was to deal with issues moreover that a human character would. Notwithstanding, after some time, thought moved to performing unequivocal undertakings, affecting deviations from science. Counterfeit neural systems have been utilized on an assortment of assignments, including PC vision, talk insistence, machine clarification, casual network isolation, playing board and PC games and accommodating finding.

4.1.7 Support vector machines (SVM)

Support vector machines are utilized to explore data used for classification and regression assessment. Given a lot of preparing models, each set apart as having a spot with both of two classes, a SVM getting ready figuring gathers a model that distributes new counsels for one class or the other, making it a non-probabilistic twofold direct classifier (paying little mind to the way that procedures, for example, Platt scaling exist to utilize SVM in a probabilistic social event setting). A SVM model is a delineation of the models as focuses in space, planned so the instances of the different classes are partitioned by a conspicuous opening that is as wide as could be typical

thinking about the current circumstance. New models are then planned into that tantamount space and anticipated to have a spot with a request dependent on which side of the initial they fall.

Following is flow chart which depicts the working of base classifiers.

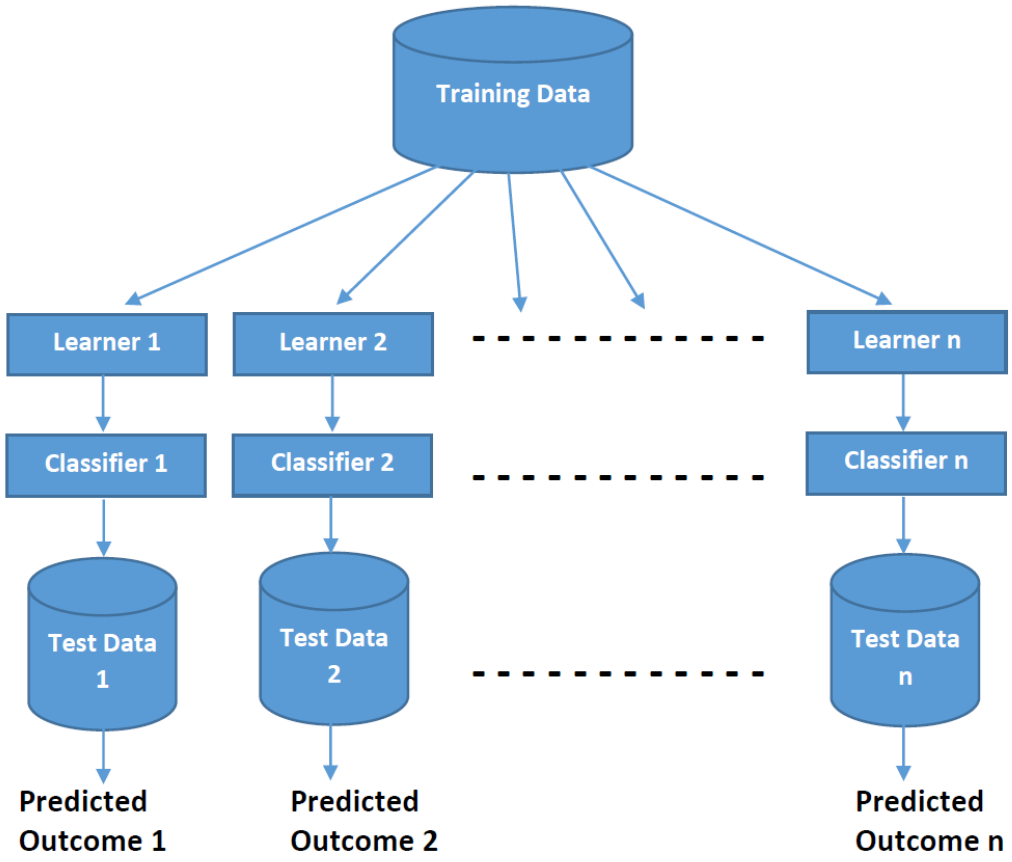

Figure 4.1 Working of Base Classifiers

The figure 4.1 clearly shows that different classifiers are learned & then tested. The predicated outcome shows accuracy of each classifier individually.

Seven base classifiers are used in this study.

We reduced the number of classifiers to only a few classes/algorithms and not different implementations of the same algorithm. but it Could be confusing that how many can be different implementations of the same algorithm if we use algorithms of same family of algorithms, and how many were variations within the same "family" of classifiers. We believe that for practitioner

and research communities, it is more useful to understand how different families of classifiers rank in relation to each other.

Following is algorithm and computational running time of the technique mentioned in above flow diagram.

4.2 Base Classifier Algorithm

---

1) Given a set of N points (samples). In this study say (N=7).
2) Convert the given set into training set (70%) and test set (30%).
3) Apply same precondition on all these datasets.
4) For i=1 to N do.
5) For j=1 to M do.
6) Apply Classifier g(i) on (Dataset(j))
7) End for
8) Save precondition [i, j]
9) End for

Here g(i) is the **ith** classifier and datasets **j** in **jth** dataset, prediction **[i, j]** is the results produced by **ith** classifier on **jth** dataset here **i=1,2, 3……...7 and j=1,2,3**.

In this case **N=7** as seven classifiers are used and **M = 3** for three different datasets.

The computational complexity of this Algorithm is O($MN \sum_{i=1}^{n} g_i(N)$).

simplifying above expression we get O **(MN** O (Max($\boldsymbol{g_i(N)}$))**.**

As line no. 4 runs N times and line 5 runs M times and since running time of different classifier is different so their complexity is described with general notation g_i (N).

4.2.1 LDA Classifier

The detail of classifier with better result i.e.**, LDA** is discussed next in this chapter.

In our observation of results, **LDA** provides best result over all datasets. Here is brief description of how LDA works; it's running time etc.

The basic concept behind **LDA** is to convert a large dimensional space to a lower dimensional space. It works on these basic steps:

1. It maximizes between class variance, **i.e.,** the distance between maximizes means of different classes.
2. It maximizes within class variance **i.e.,** distance between means and sample of each class.
3. Thirdly it constructs a lower dimensional space which maximizes the with in class variance and between class variance.

i) To calculate between class variance S_B.

To explain how above mention steps are achieved following assumptions are made.

Given in the data **X= {x_1 ,x_2 ……. x_n }** ,where $X_i$is **ith** object or sample in given data space. We have **n** observations, let there are m feature of this sample space then $X_i \epsilon R^m$ **I.e. X** is m-dimensional space. Since two class data sets is used in this study. So, X = {W_1, W_2}

Assume we have **5K** sample for each class. So, there are total **10K** samples in our dataset. We can say $\mathbf{n_1 = n_2 = 5K}$ where n_i is the number of samples of ith class and

$$\mathbf{N= \sum_{i=1}^{n} n_i = n_1 + n_2} \quad \textbf{(4)}$$

S_B is the between class variance. Which is denoted by **(m_i-m)** and is calculated as follows.

$$(m_i - \mathbf{m})^2 = \quad (W^T \mu_i \; - W^T \mu)^2 = \; WT(\mu_i \; - \mu)^T \; \mathbf{W} \quad \textbf{(5)}$$

Where $m_i = W^T \mu_i$ represents projection of mean of **ith** class and m= $W^T \mu$ represents perfection of total mean of all classes. **W** represents transformation matrix.

Here $\boldsymbol{\mu_j}$ and **μ** can be calculated as follows.

$$\mu_j \; = \frac{1}{n_j} \sum_{x_i \, \epsilon W_j} \; x_i \quad \textbf{(6)}$$

$$\mu = \frac{1}{N} \sum_{i=1}^{n} X_i \; = \quad \sum_{i=1}^{C} \frac{n_i}{N} \mu_i \quad \textbf{(7)}$$

Here **C = 2** is number of classes in our sample dataset.

The term $(\mu_i \; - \mu)(\mu_i \; - \mu)^T$ **r**epresents the between class variance of **ith** class SB_i. So, equation **(1)** can be written as

$$(m_i \; - m)^2 = W^T SB_i \, \mathbf{W} \quad \textbf{(8)}$$

Total between class variance is calculated as follows.

$$SB_i \; = \sum_{i=1}^{C} n_i \; SB_i \quad \textbf{(9)}$$

ii) Calculating with in class variance $\boldsymbol{S_W}$

Within class variance of each class $\boldsymbol{SB_j}$ is calculated as

$$\sum_{x_{i\epsilon W_j}, j=1,2,\ldots c}(\boldsymbol{W^T x_{i-}\, m_j})^2 \quad \textbf{(10)}$$

after solving we get

$$\sum_{x_{i\epsilon W_j}, j=1,2,\ldots c} \boldsymbol{W^T SW_j\, W} \quad \textbf{(11)}$$

And within class variance $\boldsymbol{SW_j}$ can calculated as:

$$\boldsymbol{SW_j} = \boldsymbol{d_j}\,\mathbf{T}^*\boldsymbol{d_j} = \sum_{i=1}^{n_j}(\boldsymbol{X_{ij}} - \mu_i)(\boldsymbol{X_{ij}} - \mu_i)^T \quad \textbf{(12)}$$

Where $\boldsymbol{X_{ij}}$ represents **ith** sample of **jth** class.

The total within class variance is

$$\mathbf{SW_j} = \sum_{i=1}^{2} \boldsymbol{Sw_i} = \sum_{x_{i\epsilon w1}}(\boldsymbol{x_i} - \mu_1)(\boldsymbol{x_i} - \mu_1)^T + \sum_{x_{i\epsilon w2}}(\boldsymbol{x_i} - \mu_2)(\boldsymbol{x_i} - \mu_2)^T \quad \textbf{(13)}$$

iii) Constructing Lower Dimensional Space

Now last step in **LDA** is construct a lower dimensional space using eigen values and eigen vectors. The transformation matrix can be calculated by using **Fisher's criterion** as follows.

$$\textbf{Arg MAX(W)}\frac{\boldsymbol{W_T S_B\, W}}{\boldsymbol{W_T S_W\, W}} \quad \textbf{(14)}$$

This formula can be reproduced and rewritten as follow:

$$\boldsymbol{S_W\, W} = \lambda \boldsymbol{S_B\, W} \quad \textbf{(15)}$$

Where λ represents the eigen values of transformation **matrix (W).** The solution of this problem can be obtained by calculating the given value $(\lambda = \lambda_1, \lambda_2 \ldots\ldots\ldots\ldots \lambda_{\boldsymbol{m}})$ and given vectors $\mathbf{V}=\{\boldsymbol{V_1}, \boldsymbol{V_2}\ldots\ldots\ldots\ldots\ldots \boldsymbol{V_m}\}$ of $\boldsymbol{S_W^{-1}\, S_B}$, If $\boldsymbol{S_B}$ is non-singular.

The Eigen vectors with **K** highest Eigen values are used to construct a lower dimensional space $(\boldsymbol{V_K})$ while other Eigen vectors $\{\boldsymbol{V_{K+1}}, \boldsymbol{V_{K+2}} \ldots\ldots\ldots\ldots \boldsymbol{V_m}\}$ are neglected.

So, the dimension of original data matrix $\mathbf{X} \epsilon \mathbf{R}^{\mathbf{N \times M}}$ is reduced by projecting it on to lower dimensional space of **LDA** $(\mathbf{V_k} \epsilon \mathbf{R}^{\mathbf{M \times K}})$.

The dimension of data is **K** after projection hence **M – K** features are ignored or deleted from each sample. This each sample $(\boldsymbol{X_i})$ which was represented as a point in **M –** dimensional space will be represented in a **K –** dimensional space$(\boldsymbol{V_K})$ as follow:

$$\boldsymbol{y_{i=}\, x_i\, V_K} \textbf{ and } \mathbf{Y}=\mathbf{X}\boldsymbol{V_K} \quad \textbf{(16)}$$

4.2.2 LDA Algorithm

1: Given a set of **N** sample$[\boldsymbol{xi}]_{\boldsymbol{i=1}}^{\boldsymbol{N}}$**,** each of which is represented as a row of length **M** and $\boldsymbol{X}$ **(N X M)** is given by,

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x(1,1)} & \boldsymbol{x(2,1)} \dots\dots\dots & \boldsymbol{x(1,M)} \\ \boldsymbol{x(2,1)} & \boldsymbol{x(2,2)} \dots\dots\dots & \boldsymbol{x(2,M)} \\ \boldsymbol{x(N,1)} & \boldsymbol{x(N,2)} \dots\dots\dots & \boldsymbol{x(N,M)} \end{bmatrix} \quad \textbf{(17)}$$

2: Computer the mean of each class $\mu_{\boldsymbol{i}}$ $(\boldsymbol{1XM})$.

3: Computer the total men of all data $\mu(\boldsymbol{1\ X\ M})$**.**

4: Computer within-class matrix $\boldsymbol{S_B}$ $(\boldsymbol{M\ X\ M})$ as follow:

$$\boldsymbol{S_B} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n_j}} \boldsymbol{n_i} (\mu_{\boldsymbol{i}} - \mu\ )(\mu_{\boldsymbol{i}} - \mu\ )^{\boldsymbol{T}} \quad \textbf{(18)}$$

5: Compute within-class matrix $\boldsymbol{S_w}(\boldsymbol{M\ X\ M})$as follows.

$$\boldsymbol{S_B} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{c}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n_j}} (\boldsymbol{x_{ij}} - \mu_{\boldsymbol{j}}\ )(\boldsymbol{x_{ij}} - \mu_{\boldsymbol{j}})^{\boldsymbol{T}} \quad \textbf{(19)}$$

Where $\boldsymbol{x_{ij}}$ represents the ith sample in the **jth** class.

6: the matrix W that maximizing **Fisher's formula** is calculated as follow,

$$\mathbf{W}=\boldsymbol{S_W^{-1}S_B} \quad \textbf{(20)}$$

The eigenvalues **(ƛ)** and eigenvectors **(V)** of **W** are then calculated.

7: Sorting eigenvectors in descending order according to their corresponding eigenvalues. The first **k** eigenvectors are then used as a lower dimensional space ($\boldsymbol{V_K}$).

8: Project all original sample **(X)** onto the lower dimensional space of **LDA**.

4.2.3 kNN Classifier

Results shown below in this chapter confirms that performance of kNN classifier is below par among base classifiers. Here we will try to discuss reasons that why kNN's performance is not up to the mark on this particular kind of data.

We will also be discussing its algorithm and working of kNN algorithm.

KNN is a very simple algorithm used to solve classification problems. KNN stands for K-Nearest Neighbors. K is the number of neighbors in KNN.

kNN is known as lazy learner or instance-based learner. It has no training time as one has not to learn anything in training period. It stores training examples and learn from it while testing or making predictions. It is very easy to implement as it uses only two parameters i.e. k and distance function.

Different distance functions can be used in kNN e.g.

Makowskis	$\sum_{i=1}^{k}[(\lvert x_i - y_i \rvert)^q]^{1/q}$
Euclidian	$\sum_{i=1}^{k}[(x_i - y_i)^2]^{1/2}$
Manhattan	$\sum_{i=1}^{k} \lvert x_i - y_i \rvert$
Hamming	$d = min\,\{d(x,y): x,y \in C, x \neq y\}$

With all these benefits of using kNN, it did not work well with the data we selected. Following are the reasons for not performing well.

kNN Does not work well with large dataset. Since we used quite a large data set in all three examples, so it did not perform well. The cost of calculating the distance between the new point and each existing point is huge which degrades the performance of the algorithm. Also, high dimensionality of our data sets is another reason behind bad performance of kNN. Since our data consist of reviews of customers, so it may have missing values, noisy data and outliers. Which can be a big cause in the bad performance of kNN.

kNN works by analogy that whom do you resemble the most. i.e. if a test sample is provided, and you find the k nearest neighbors of that sample in training examples. And by examining plurality vote you predict the class of that test sample

Following figures illustrates the concept behind kNN process.

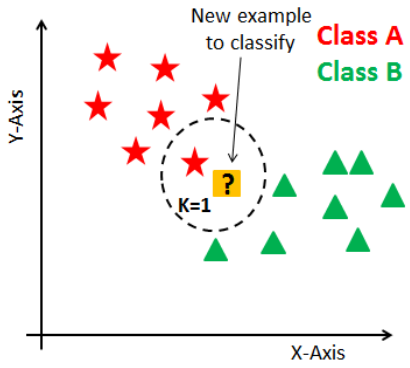

Figure 4.2 illustrates the concept behind kNN with k=1

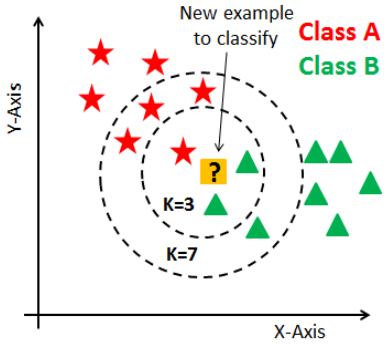

Figure 4.3 illustrates the concept behind kNN with k=3

In figure 4.2 above it is clear that if k=1, the nearest neighbor will be class of query object. While in figure 4.3 one can see that changing value of k can affect the predicted class. Since kNN is based on feature similarity so value of k can be critical. Choosing right value of k is a process called parameter tuning. Usually square root of n (where n is number of data samples) is good chosen value of n, odd value is normally taken to avoid any confusion.

Following is flow diagram of kNN process.

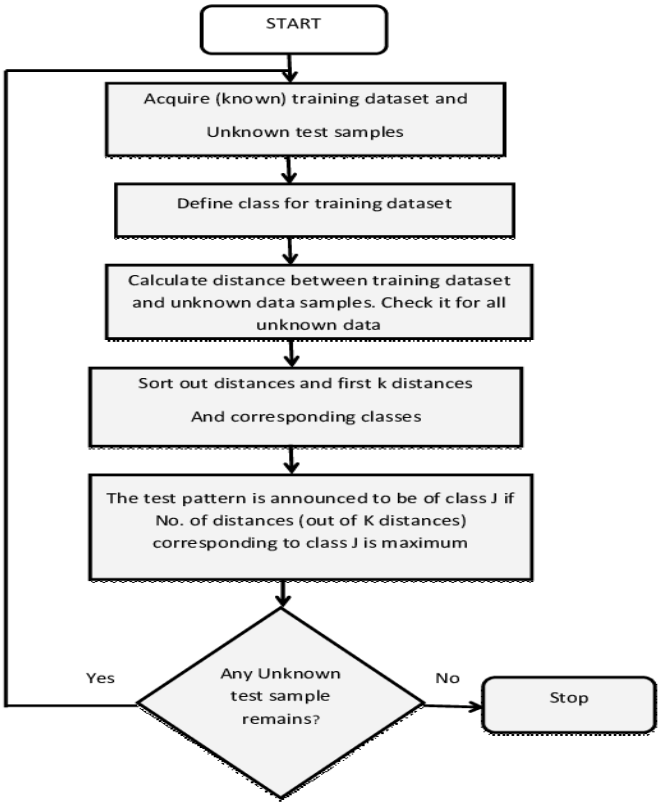

Figure 4.4 flow diagram of kNN process

Above process can be described mathematically as follows.

Let $\mathbf{x} \in \mathbf{D}$ and $\mathbf{D} = \cup_{i=1}^{n} xi$, $\mathbf{i = 1 \ldots n}$, is the data set we are using. And let $\mathbf{y} \in \{1, 2, \ldots t\}$. Since we are dealing with two class problem, so $\mathbf{y} \in \{0,1\}$.

Consider the target function $\mathbf{f(x) = y}$. i.e. f assigns y (0 or 1) to every value of $\mathbf{x} \in \mathbf{D}$.

so, $\mathbf{f: D \rightarrow \{0,1\}}$. Now if we found k nearest neighbors $\mathbf{D_k \subset D}$ to a test sample $\mathbf{x_q}$. We can say $\mathbf{D_k = \{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_k, f(x_k)\}}$.

We can define kNN hypothesis as $\mathbf{Arg \max_{y \in \{0,1\}} \sum_{i=1}^{k} \delta(y, f(xi))}$. Here δ denotes delta function

$$\delta(a,b) = \begin{cases} 1 \; if \; a = b \\ 0 \; if \; a \neq b \end{cases} \qquad (21)$$

Or we can simply say $\mathbf{h(x_t) = mode \ (\{f(x_1), (f(x_2), \ldots, (f(x_k)\})}$.

Following is algorithm for kNN classifier.

4.2.4 kNN Algorithm

---

kNN(Q,T,k,d)

//Query sample $\mathbf{x_1, x_2, \ldots, x_m}$

//Training sample $\mathbf{x_1, x_2, \ldots, x_n}$

//k: parameter used for nearest neighbor.

//d: Number of features for each data sample.

Initialization (Load Data)

For i= 1 to n

For j=1 to m

calculate$(\boldsymbol{x_i}, \boldsymbol{x_j}) = (\sum_{t=1}^{d}(\mathbf{A}_t(\mathbf{x}_i) - \mathbf{A}_t(\mathbf{x}_j))^2)^{1/2}$

Save $D(\boldsymbol{x_i}, \boldsymbol{x_j})$

Sort the above saved collection in ascending order

Pick first k entries

Get labels (Class value) of selected k entries.

Return mode of k labels

The working of this algorithm has been discussed earlier. The computational complexity of this algorithm depends upon the number of query samples, training samples and number of features. For one query sample and very few numbers of features, the running time of kNN will be **O(n)**. Where 'n' is number of training samples. But to generalize it for larger number of query samples and significant number of features, the running time will be **O(dmn)**. Where 'd' is number of features and 'm' is number of query data samples.

4.3 Notations Used.

Table 4-2 Notations Used

Notation used	Description
X	Data matrix
N	Total number of attributes in X
K	Total number of samples in X
n_j	Number of samples in $\acute{\omega}_i$
μ_i	The mean of the ith class
Sw_i	Total or global mean of all samples
SB_i	Within-class variance or scatter matrix of the ith class ($\acute{\omega}_i$)

V	Between-class variance of the ith class ($\acute{\omega}_i$)
K	The **ith** sample in the jth class.
X_i	The dimension of the lower dimensional space (V_k)
M	Dimension of X or the number of features of X
V_i	Eigenvector of **W**
X_{ij}	ith eigenvector
V_k	The lower dimensional space
C	Total number of classes
m_i	The mean of the **ith** class after projection
M	The total mean of all classes after projection
S_w	Within-class variance
S_B	Between –class variance
$\acute{\lambda}$	Eigenvalue matrix
$\acute{\lambda}_i$	**Ith** eigenvalue
Y	Projection of the original data
$\acute{\omega}_i$	**Ith** class

4.4 Evaluation metric used:

4.4.1 Confusion Matrix:

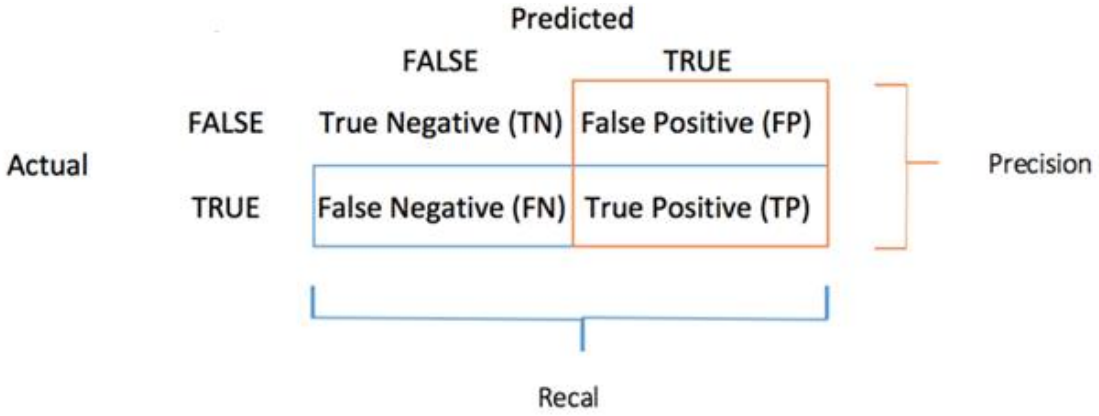

Figure 4.5 Confusion Matrix

In this study accuracy measure and ROC curve are used to represent the results of different classifiers and to show comparison between them.

4.4.2 Accuracy

Accuracy measure tells us about the correct predictions made by a classifier. It is % age of correctly measured true and false class values out of total class values.

$$\text{Accuracy} = \frac{TP+TN}{P+N} \qquad \textbf{(22)}$$

4.4.3 Receiver Operating Characteristic (ROC) Curve

Receiver Operating Characteristic or ROC Curve is graph of TPR (True positive Rate) vs FPR (False positive Rate) for different cut off values between 0 & 1. FPR is taken along x-axis and TPR is taken along y-axis. ROC curve shows the diagnostic ability of classifier. ROC space consists of all prediction instances of confusion matrix.

TPR = Sensitivity measure (True positive rate)

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad \textbf{(23)}$$

FPR = 1 - Specificity measure (True negative rate)

$$\text{Specificity} = \frac{TN}{FP+TN} \qquad \textbf{(24)}$$

Following figure clearly elaborates the behavior of classifiers while using ROC curve.

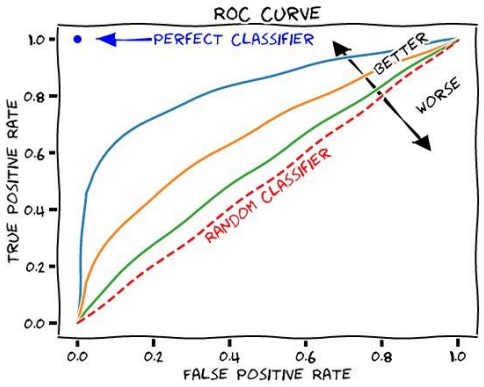

Figure 4.6 elaboration of ROC

Figure clearly elaborates the performance of different classifiers. The classifier represented with green curve shows worst performance as it is very close to diagonal, so area under curve with respect to diagonal is very less as compared to curve in light blue. So, the curve represented in light blue shows best performance in this particular case. A perfect classifier will have cot off

points close to (0,1) i.e. top left corner of graph. While a point on diagonal say (0.5,0.5) does not provide any information as it shows 50% accuracy. Which is worthless for any prediction.

Area under curve (AUC) with respect to diagonal is another accuracy measure. It is used to summarize accuracy of a random classifier measure with single value. If value is 1, the classifier is ideal one and it shows perfect test while the classifier with 0.5 value is considered as worst. Usually classifiers with values more than 0.7 to 0.8 are taken as acceptable classifiers. The classifiers with values ranging between 0.8 to 0.9 are considered excellent. While the classifiers with values more than 0.9 are considered outstanding.

Different classifiers can be compared on a single ROC curve graph and their performance can be measured by AUC (area under curve).

the area under the curve is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

4.5 Results obtained from base classifiers:

4.5.1 Base classifiers with all datasets of 10000 instances:

Figure 4.7 shows the results of all algorithms over Amazon Reviews Dataset with 10000 instances. The three algorithms i.e. SVM, LDA and NNET outperformed other algorithms with the accuracy of 80%, whereas NB, CTREE, and KNN got the accuracy of 76%, 74% and 68% respectively. The RPart performed worst in this case with the accuracy of 65% only.

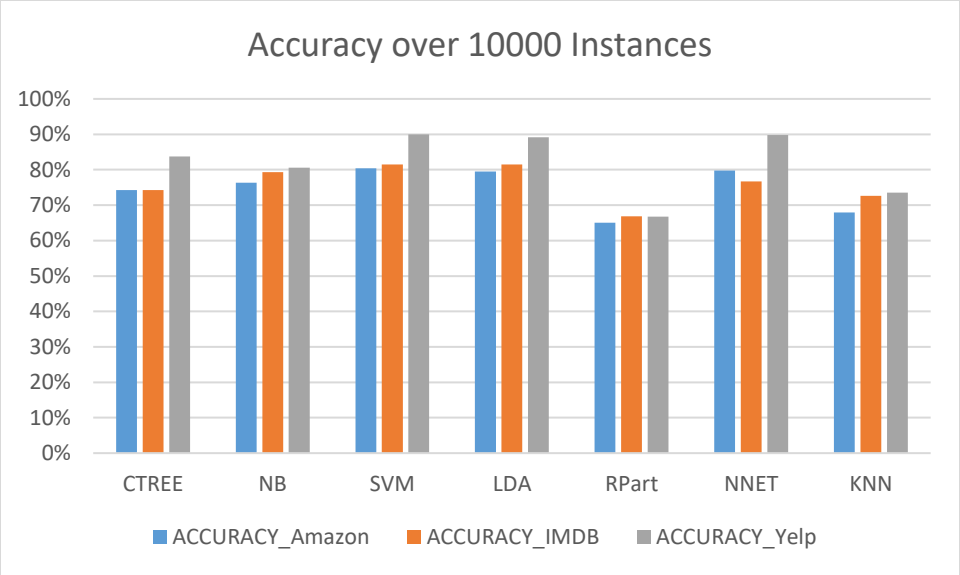

Figure 4.6 Accuracies of all classifiers over Selected Dataset with 10000 instances

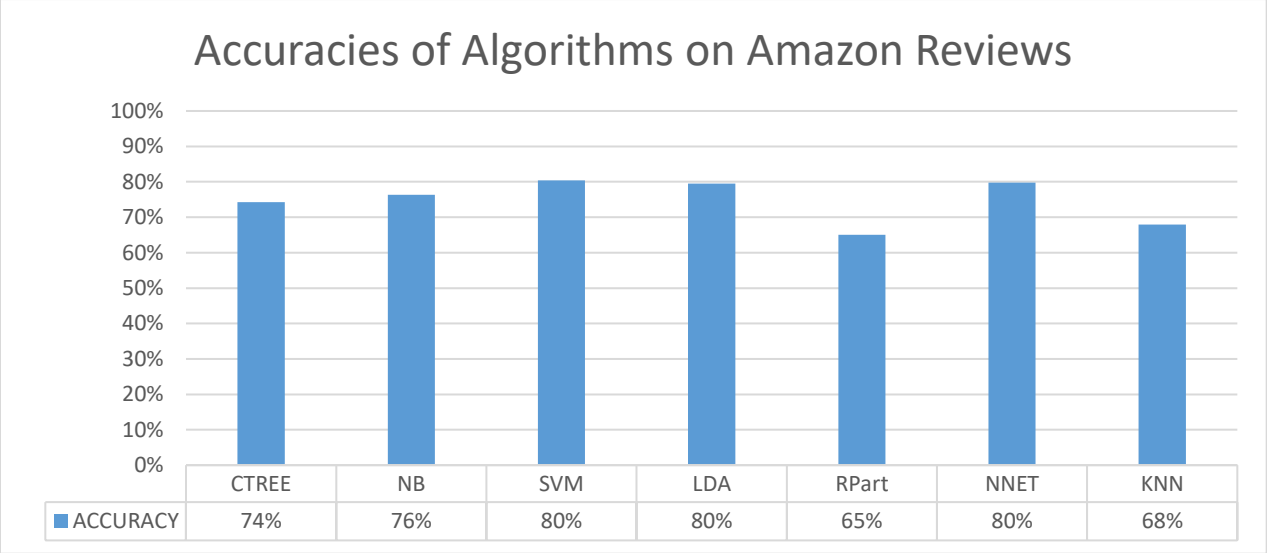

Figure 4.7 Accuracies of all classifiers over Amazon Reviews Dataset with 10000 instances

Figure 4.8 shows the results of all classifiers over IMDB Movie Reviews Dataset with 10000 instances. The two classifiers i.e. SVM and LDA outperformed other classifiers with the accuracy of 82%, whereas NB, NNET, CTREE, and KNN got the accuracy of 79%, 77%, 74% and 73% respectively. The RPart performed worst in this case with the accuracy of 67% only but improved as compared to the performance in case of Amazon Reviews Dataset.

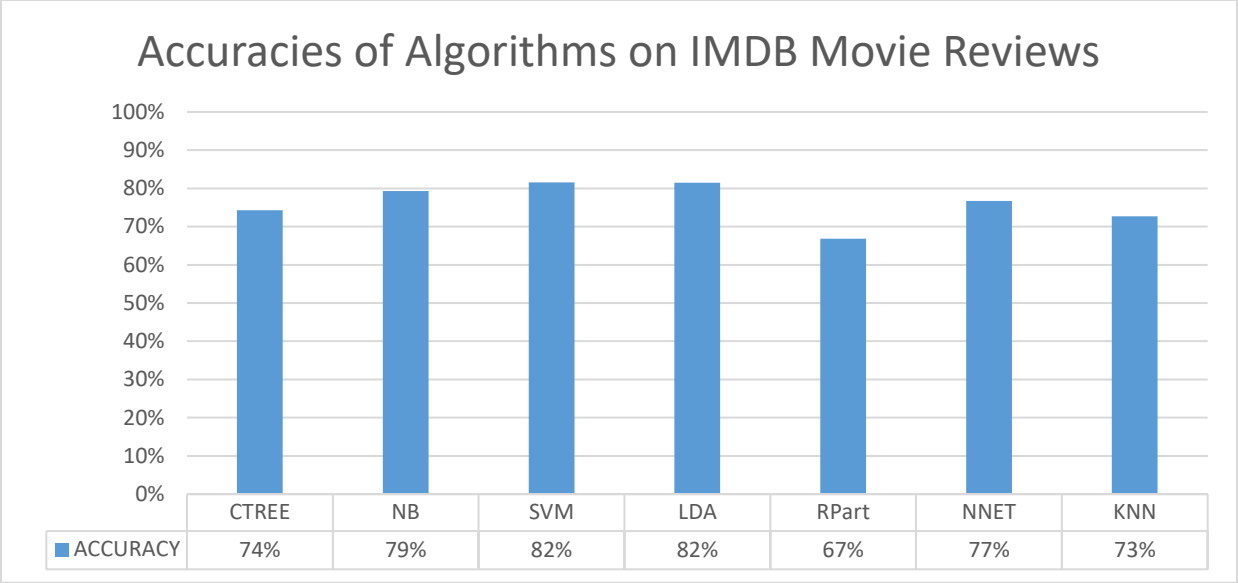

Figure 4.8 Accuracies of all classifiers over IMDb Dataset with 10000 instances

Figure 4.9 shows the results of all classifiers over YELP Dataset with 10000 instances. The classifiers performed better than previous two datasets i.e. SVM and NNET outperformed other classifiers with the accuracies of 90%, whereas LDA, CTREE, NB, and KNN got the accuracy of 89%, 84%, 81% and 74% respectively. The RPart once again performed worst with the accuracy of 67% only but improved as compared to the performance in case of Amazon Reviews Dataset.

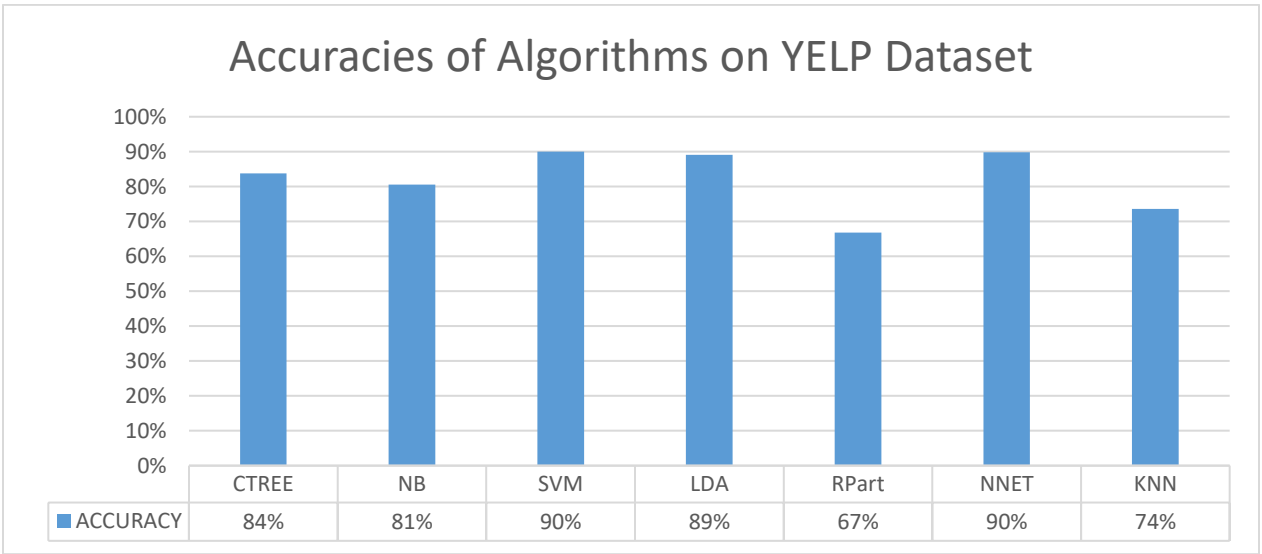

Figure 4.9 Accuracies of all classifiers over Yelp Dataset with 10000 instances

4.5.2 Base Classifiers with all datasets of 6000 instances

It can be seen from figure 4.11 that results have been decreased slightly as Amazon Reviews dataset sizes reduced from 10000 instances to 6000 instances. Figure 4.4 shows that LDA

outperformed all other classifiers with an accuracy of 79% whereas RPart again performed worst with the accuracy of 63%.

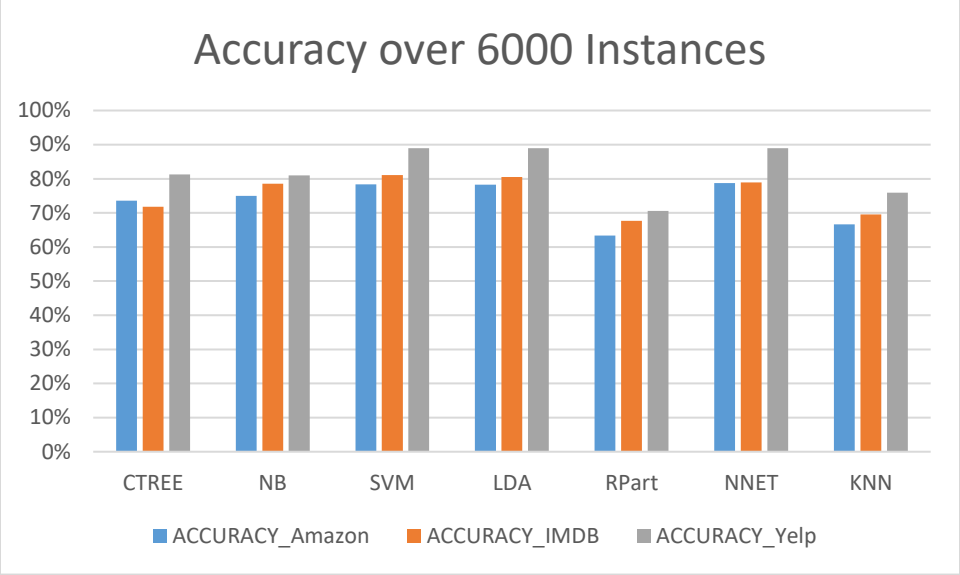

Figure 4.10 Accuracies of all classifiers over All Datasets with 6000 instances

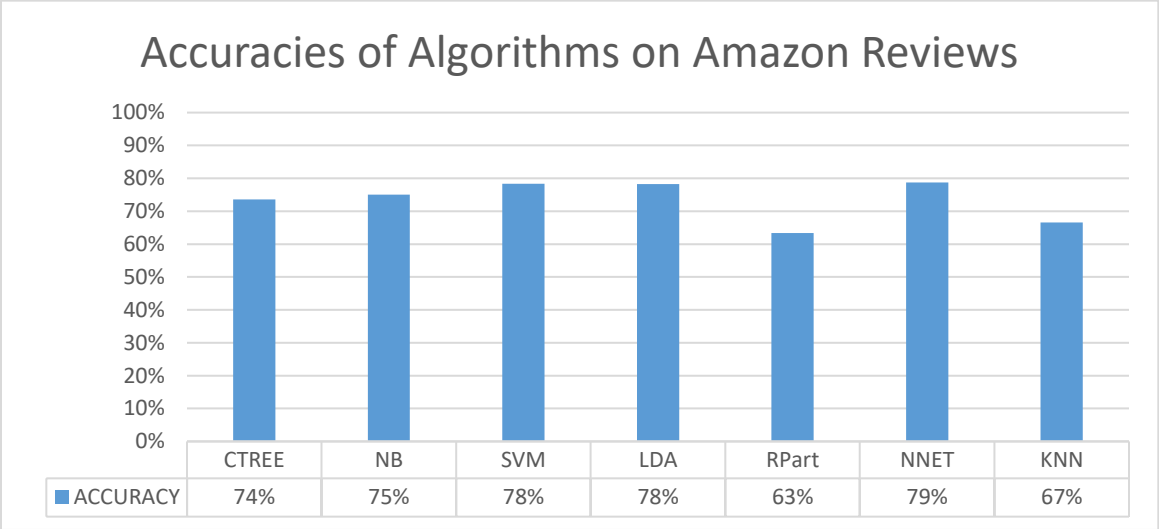

Figure 4.11 Accuracies of all classifiers over Amazon Reviews Dataset with 6000 instances

Following figure shows that in more of the cases results have been decreased slightly as IMDB Reviews dataset sizes reduced from 10000 instances to 6000 instances. It also shows that LDA and SVM outperformed all other classifiers with an accuracy of 81% whereas RPart again performed worst but better than all previous cases with the accuracy of 63%.

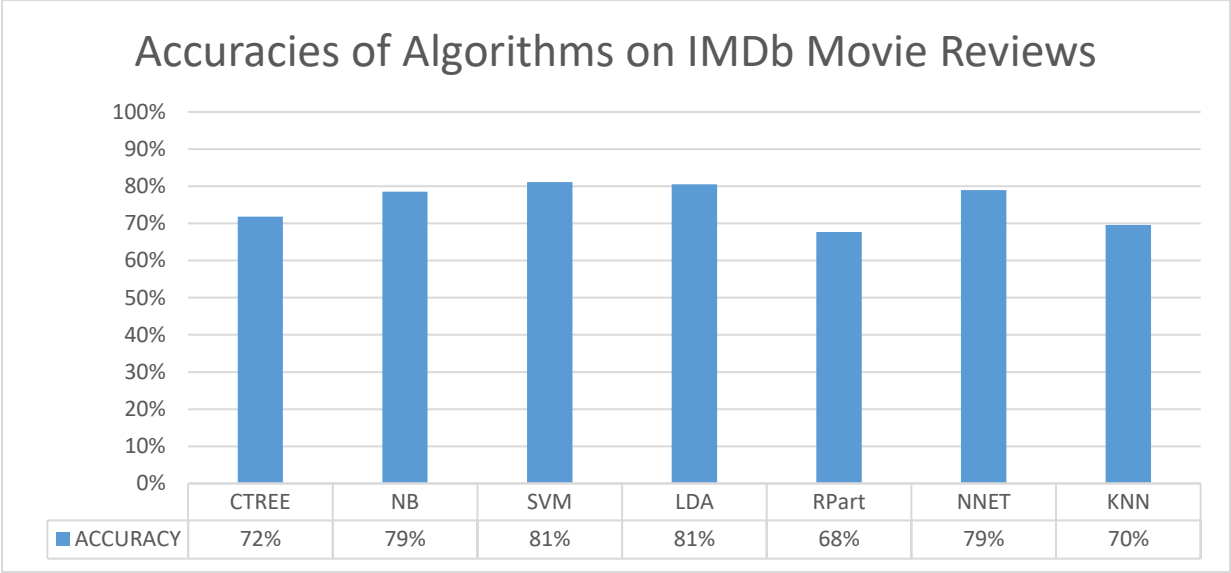

Figure 4.12 Accuracies of all classifiers over IMDb Dataset with 6000 instances

Figure 4.13 shows the results of all classifiers over YELP Dataset with 6000 instances. It can be seen that the classifiers performed better than previous two datasets i.e. SVM, LDA and NNET outperformed other classifiers with the accuracies of 89%, whereas LDA, CTREE, NB, and KNN got the accuracy of 89%, 84%, 81% and 74% respectively. The RPart once again performed worst with the accuracy of 67% only but improved as compared to the performance in case of Amazon Reviews Dataset.

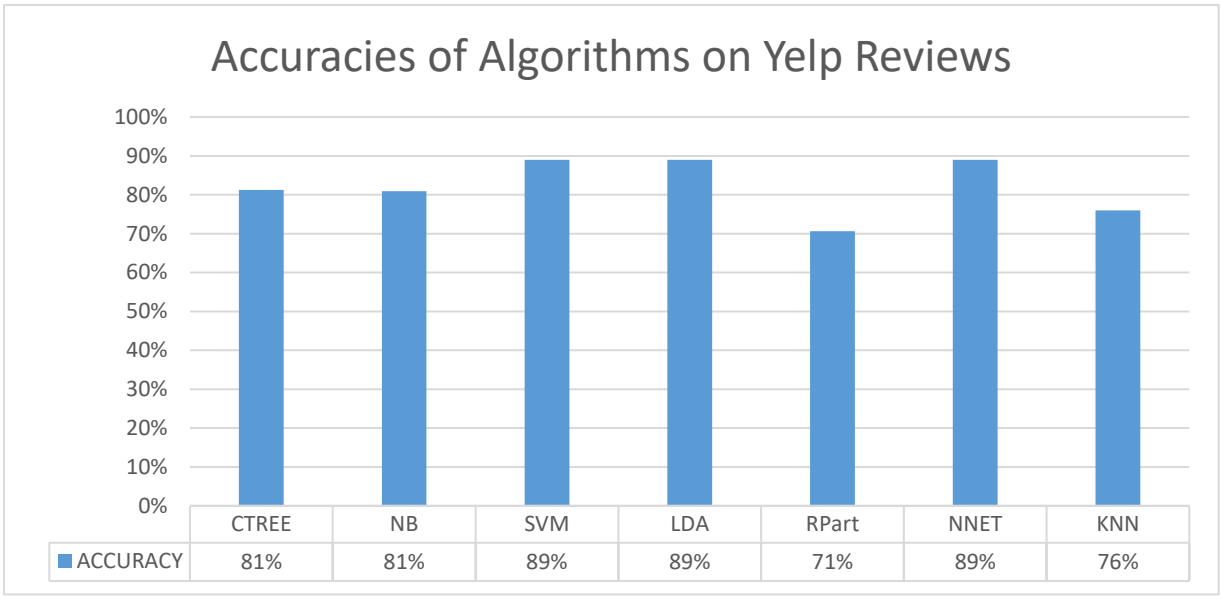

Figure 4.13 Accuracies of all classifiers over YELP Dataset with 6000 instances

4.5.3 Base Classifiers with all datasets of 3000 instances

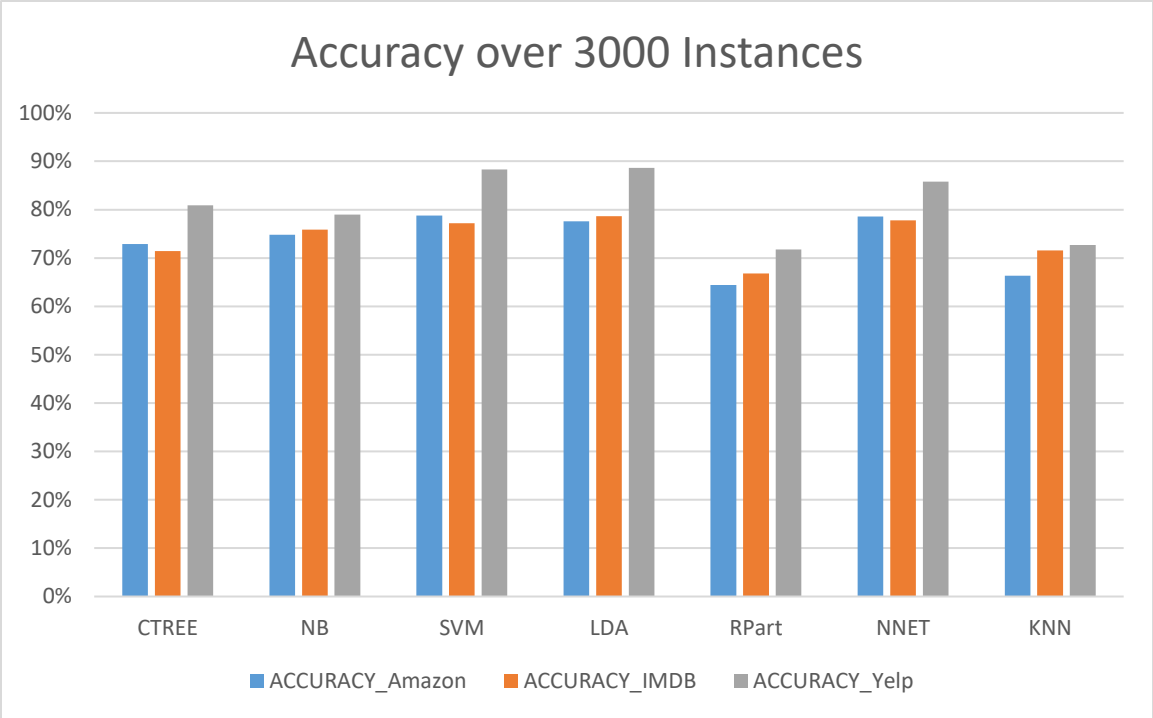

Figure 4.14 Accuracies of all classifiers over All Datasets with 3000 instances

Figure 4.15 shows the results of all classifiers over Amazon Reviews Dataset with 3000 instances. It can be seen that SVM and NNET again overtook other classifiers with the accuracies of 79%, whereas LDA, NB, CTREE, and KNN got the accuracy of 78%, 75%, 73% and 66% respectively. The RPart another time performed worst with the accuracy of 64% only.

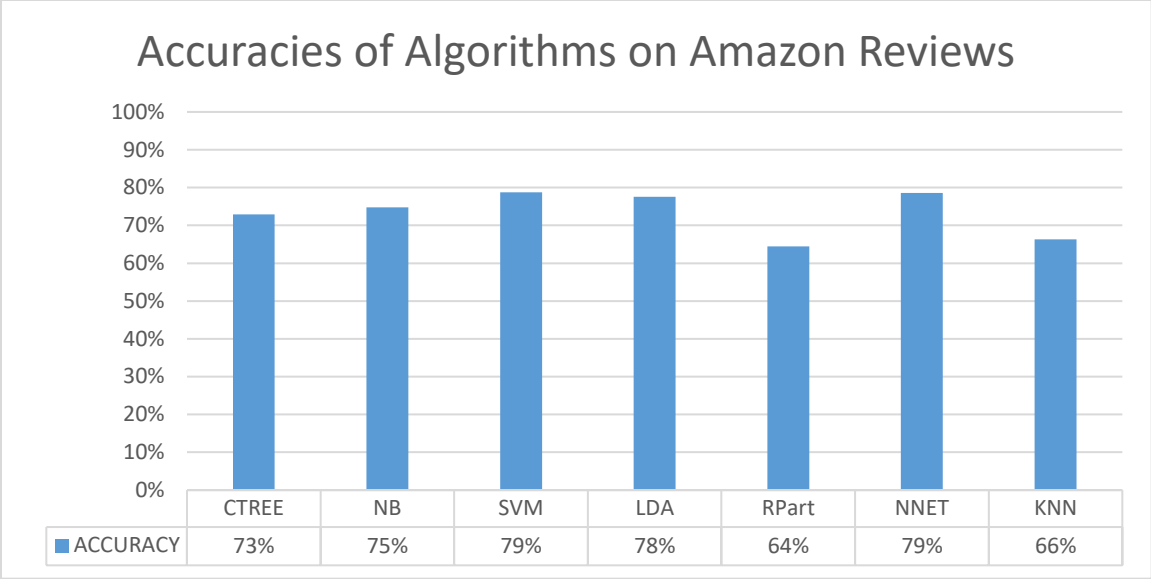

Figure 4.15 Accuracies of all classifiers on Amazon Reviews Dataset with 3000 instances

Figure 4.16 shows the results of all classifiers over IMDB Movie Reviews Dataset with 3000 instances. It can be seen that LDA in this case again overtook other classifiers with the accuracies of 79%, whereas NNET, SVM, NB, KNN, CTREE and RPart got the accuracy of 78%, 77%, 76%, 72%, 71% and 67% respectively.

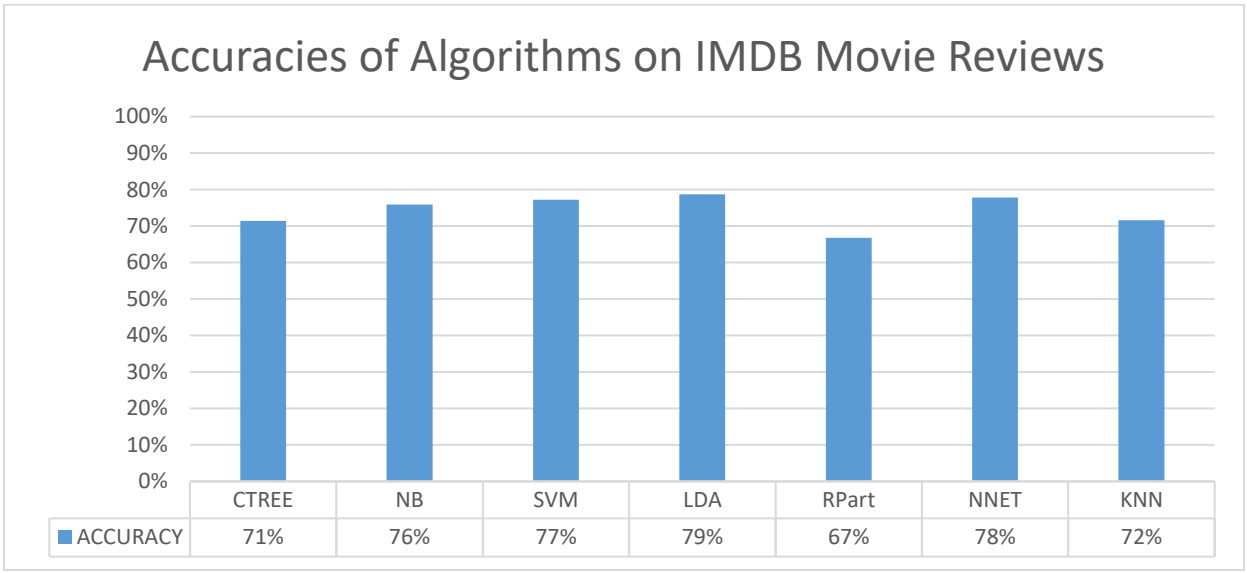

Figure 4.16 Accuracies of all classifiers on IMDb Dataset with 3000 instances

Figure 4.17 shows the results of all classifiers over YELP Reviews Dataset with 3000 instances. It can be seen that LDA in this case again overtook other classifiers with far better than IMDB and Amazon Reviews Dataset having accuracy of 89%, whereas SVM, NNET, CTREE, KNN and

RPart got the accuracy of 88%, 86%, 81%, 73% and 72% respectively. NB performed worst in this case with the accuracy of only 51%.

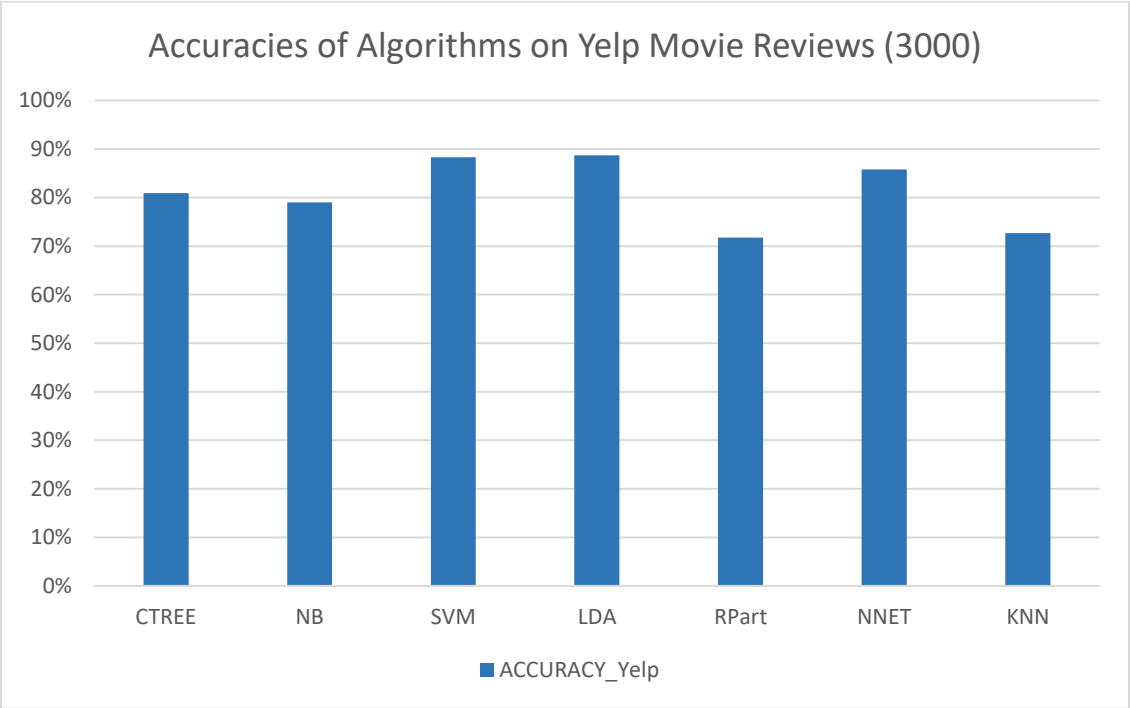

Figure 4.17 Accuracies of all classifiers on YELP Dataset with 3000 instances

4.5.4 Base Classifiers with all data sets of 1500 instances

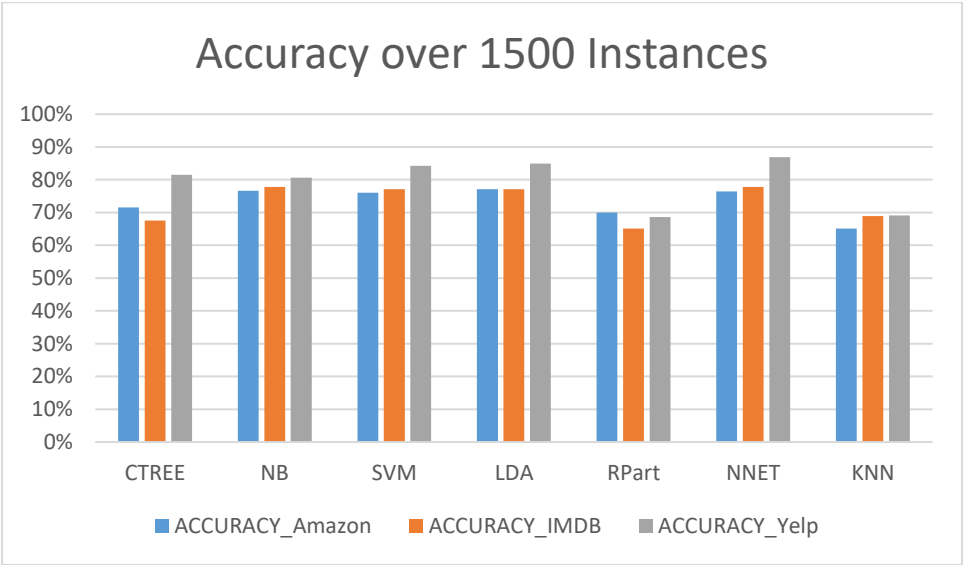

Figure 4.18 Accuracies of all classifiers on All Datasets with 1500 instances

Figure 4.19 shows the results of all classifiers over Amazon Reviews Dataset with 1500 instances. NB and LDA in this case overtook other classifiers with the accuracy of 77%, whereas SVM, NNET, CTREE, RPart and KNN got the accuracy of 76%, 76%, 72%, 70% and 65% respectively. KNN performed worst in this case.

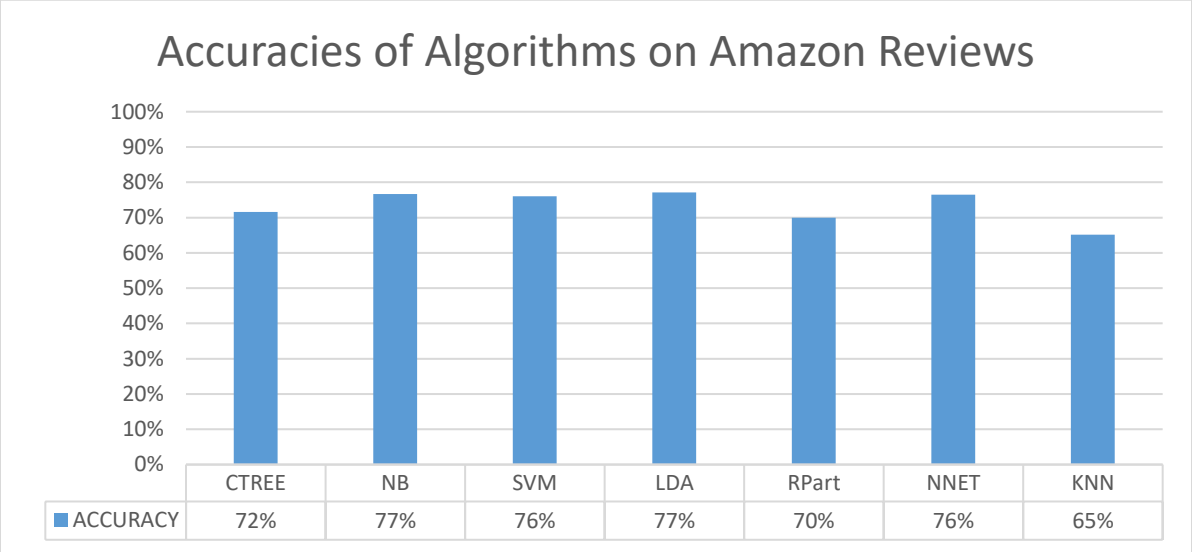

Figure 4.19 Accuracies of all classifiers on Amazon Reviews Dataset with 1500 instances

Figure 4.20 shows the results of all classifiers over IMDB Movie Reviews Dataset with 1500 instances. NB and NNET in this case outperformed other classifiers with the accuracy of 78%, whereas SVM, LDA, KNN, CTREE and RPart got the accuracy of 77%, 77%, 69%, 68% and 65% respectively. KNN performed worst in this case.

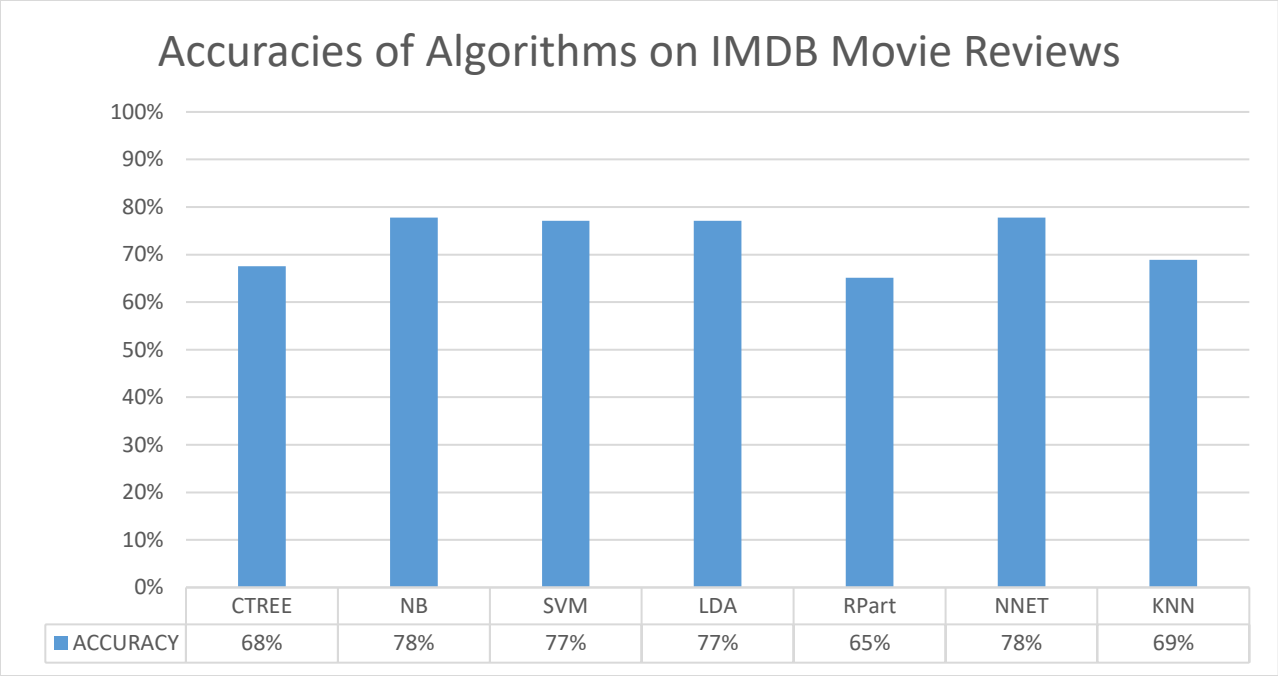

Figure 4.20 Accuracies of all classifiers on IMDB Dataset with 1500 instances

Figure 4.21 shows the results of all classifiers over YELP Reviews Dataset with 1500 instances. NNET in this case outperformed other classifiers with the accuracy of 87%, whereas LDA, SVM, CTREE, NB, KNN and RPart got the accuracy of 85%, 84%, 82%, 81%, 69% and 69% respectively.

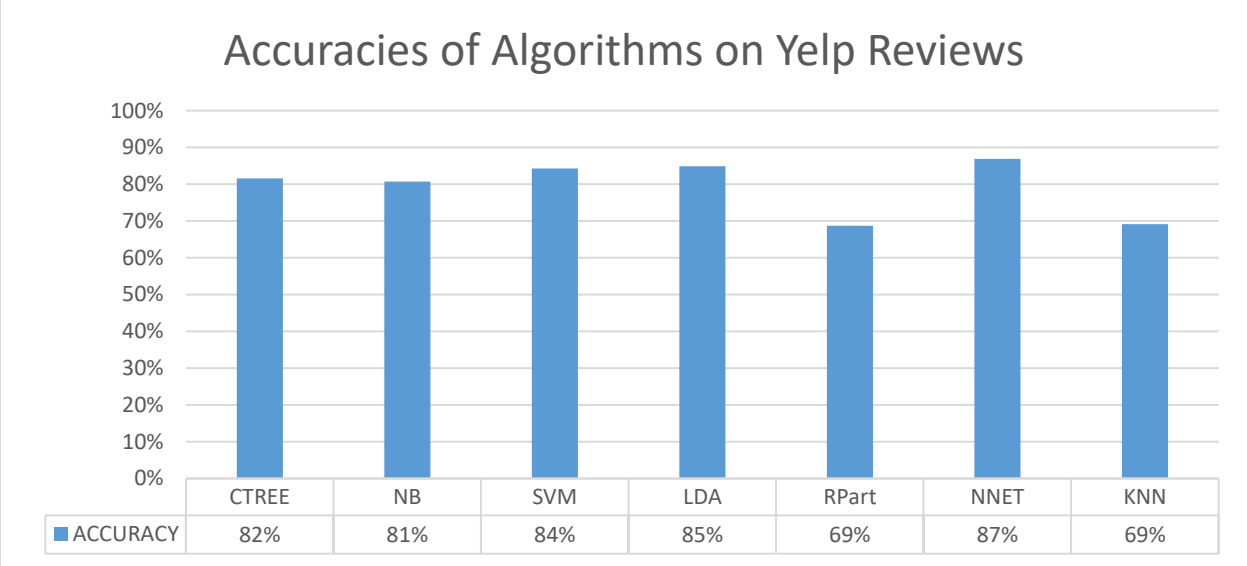

Figure 4.21 Accuracies of all classifiers on YELP Dataset with 1500 instances

Following are combined results of Base classifiers on different data sets. Results are rounded up and are in percentages.

4.6 Comparison of Base Classifiers:

Different comparison tables of results of all individual classifiers are given below. Three different measurement tools are used. All the numbers are in percentages. Acc is used for accuracy measure, Sen for sensitivity measure and Spec for specificity measure.

4.6.1 Accuracy measure of Base classifiers on Amazon data set.

Table 4-3 Accuracy of Base classifiers on Amazon data set

Classifier	**10,000 instances**			**6,000 instances**			**3,000 instances**			**1500 instances**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	74	80	68	74	74	73	73	70	76	72	70	73
NB	76	72	80	75	71	79	75	76	74	77	68	85
SVM	**80**	81	79	78	77	80	**79**	79	78	76	76	76
LDA	**80**	80	80	**78**	77	80	**78**	78	77	**77**	78	76
RPART	65	95	35	63	29	97	64	95	34	70	87	53
NNET	**80**	80	80	**79**	76	81	**79**	78	79	76	74	79
KNN	68	62	74	67	63	71	66	66	81	65	54	76

Although LDA perform consistently well on different sample sizes of Amazon data set. But still consistency is an issue with individual classifiers. In table 4-3, it is clear that at certain sample size, some other classifier is performing better than LDA. For instance, with 6000 sample size, NNET is performing better than LDA with accuracy measure of 79%. Also, SVM and NNET shows better performance with 3000 sample size with 79% accuracy.

4.6.2 Accuracy measure of base classifiers on IMDB data set.

Table 4-4 Accuracy of base classifiers on IMDB data set

Classifier	**10,000 instances**			**6,000 instances**			**3,000 instances**			**1500 instances**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	74	71	78	72	70	73	71	67	76	68	51	84
NB	79	76	82	79	76	81	76	80	72	**78**	83	73
SVM	**82**	79	84	**81**	80	83	77	76	78	77	78	76
LDA	**82**	78	85	**81**	79	82	**79**	78	80	**77**	76	78

RPART	67	46	88	68	48	87	67	44	89	65	44	86
NNET	77	77	77	79	79	79	78	71	84	**78**	77	78
KNN	73	69	79	70	50	88	72	73	70	69	64	73

Table 4-4 shows that LDA's performance is better than other classifiers. But again with 1500 sample size, NNET and NB has better accuracy with 78%. Which again questions the consistency issue.

4.6.3 Accuracy measure of base classifiers on YELP data set.

Table 4-5 Accuracy of base classifiers on YELP data set

Classifier	**10,000 instances**			**6,000 instances**			**3,000 instances**			**1500 instances**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	84	82	85	81	84	79	81	79	82	82	92	71
NB	81	72	89	81	70	91	51	0.2	99	81	74	88
SVM	**90**	89	91	89	89	89	88	89	88	84	84	85
LDA	**89**	89	89	**89**	89	89	**89**	88	89	**85**	82	88
RPART	67	38	96	71	92	49	72	93	51	69	40	97
NNET	**90**	90	90	89	89	90	86	85	86	**87**	86	88
KNN	74	85	62	76	84	68	73	78	68	69	42	96

Results on accuracy measure on different sample sizes of YELP data set also highlight the same issue of consistency. It can be seen in Table 4-5 that with 10,000 instances SVM and NNET has better accuracy than other classifiers. While with 6,000 samples LDA and SVM perform better with accuracy measure of 89%. LDA best perform with 3,000 sample size with accuracy of 89%. And NNET outperform other classifiers with an accuracy of 87%.

Now if we talk about consistency in accuracy, it can be observed from above tables that LDA secured top position on 3 occasions out 12 data samples. Its consistency rate is 25%. While SVM performed better 5 out of 12 times so its consistency rate is 42%. NNET has consistency rate of 33% as it performed best 4 times out of 12.

4.7 ROC (Receiver Operating Characteristics) curves and area under curve:

In Machine Learning, an AUC - ROC Curve is a reliable performance measurement tool. When we need to check a classification problem, we use AUC (**Area Under the Curve**) ROC (**Receiver Operating Characteristics**) curve.

Following figures explains ROC curves of two different base classifiers. One is LDA, which performed well while other is kNN, which did not perform so well.

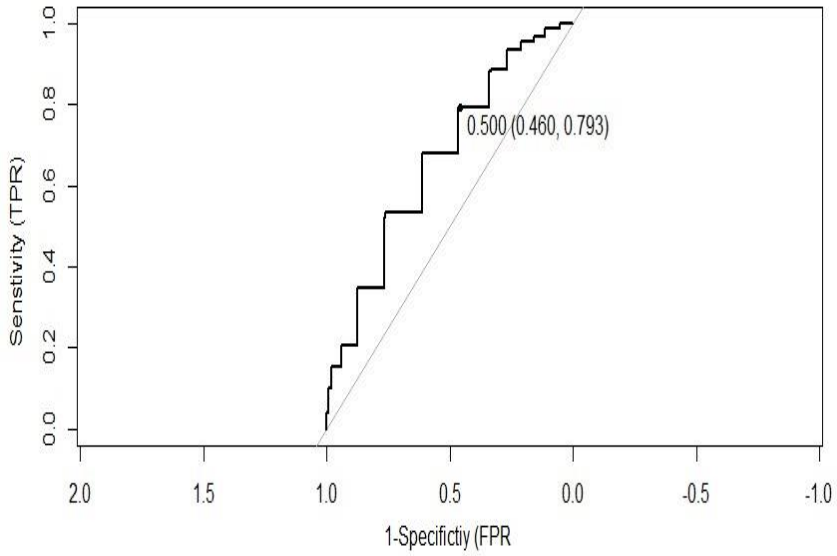

Figure 4.22 ROC for knn model

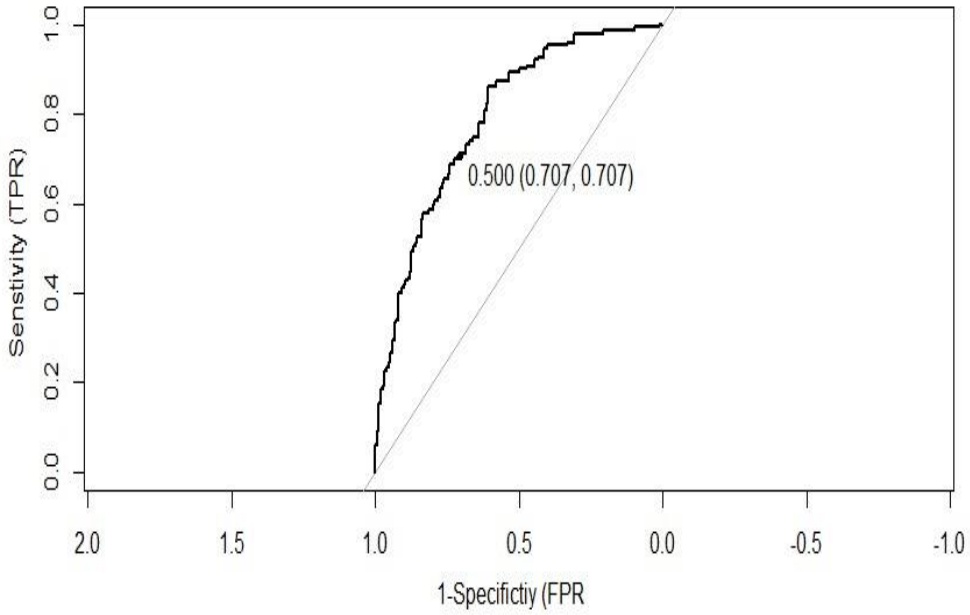

Figure 4.23 ROC for lda model

The following figure is a comparison between ROC curves of knn and lda.

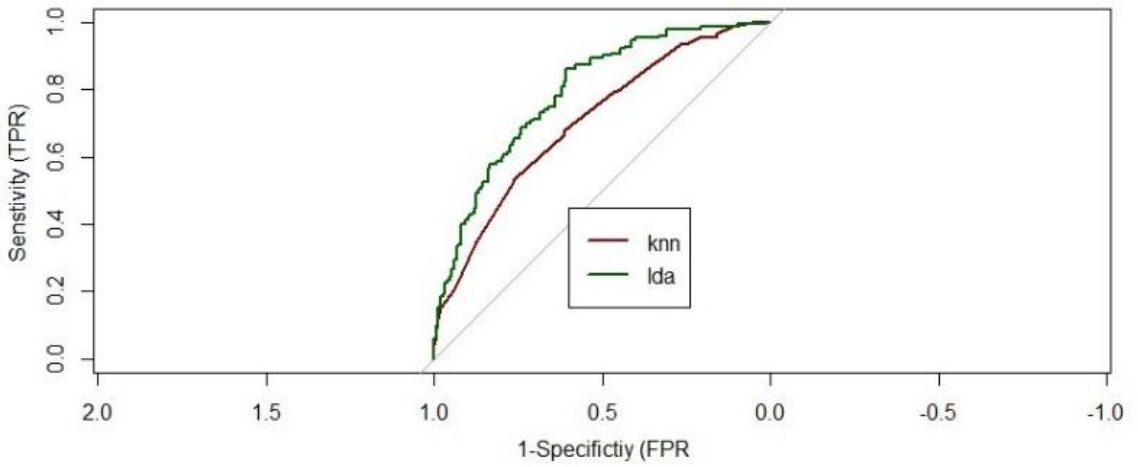

Figure 4.24 Comparison between ROC's of knn and lda

It is clearly seen that LDA has more probability of finding of getting accurate prediction on a random sample. Following figure shows the area under ROC curves.

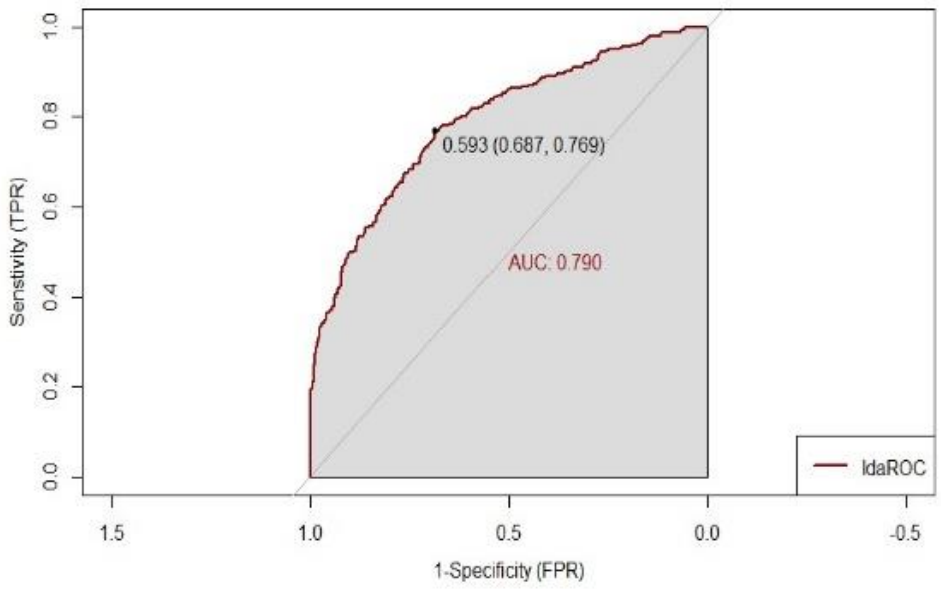

Figure 4.25 ROC and AUC of knn model

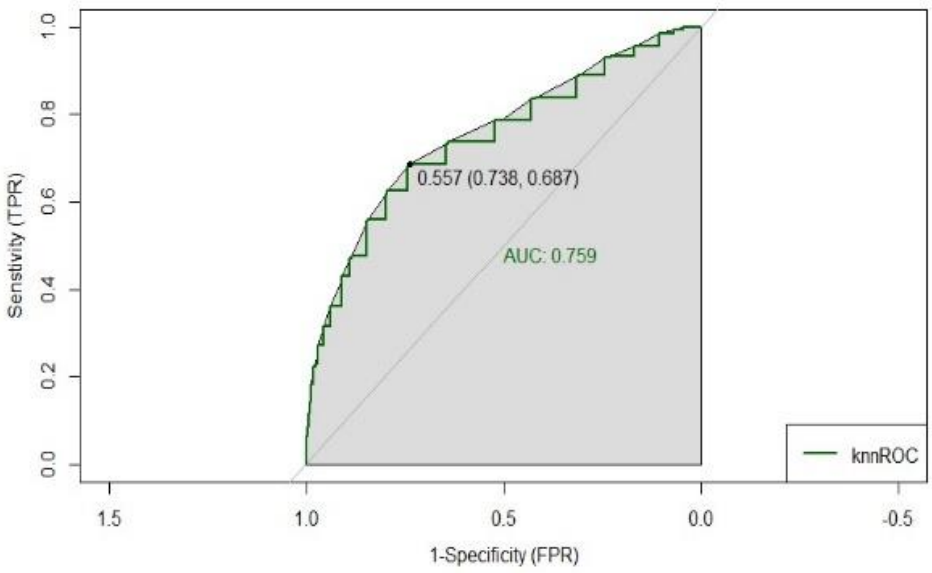

Figure 4.26 ROC and AUC of lda model

At base level seven different classifiers were used. Area under curve (AUC) of two classifiers is mentioned in above figures. AUC for lda is 0.79 which better than kNN with AUC 0.759. It shows that LDA has more chances of predicting correctly than kNN.

4.8 Discussion:

Different data samples were taken to check whether size of sample has any impact on accuracy of classifiers. It is clear from above tables that size of sample does have an impact on accuracy of individual classifier. The sequence of results varies when certain classifiers are applied if data samples are changed. Even though, LDA and SVM is more consistent than other classifiers, yet some other classifiers perform better with certain sample sizes

It can be noticed that LDA and SVM outperformed other classifiers in predicting the right class in most of cases. While KNN and RPART are not performing well. Also, KNN and RPART are slow learners. In this particular case, since number of features formed are huge, so it takes lot of time to create so many trees in RPART and find closets distances in case of KNN.

On the other hand, LDA constructs a lower dimensional space which maximizes the “within” class variance and “between class” variance and thus is able to produce better results.

Considering all three data sets, it can be concluded that individual classifiers are not only inconsistent on different data sets but also, are inconsistent on different sample sizes of same data set. So, any single classifier ca not be stated as clear winner.

This is the base of this study. Now the main goal is to find a way to improve the accuracy consistently of predicted results of above-mentioned text data with two classes.

In this chapter, different heterogenous base level classifiers are applied on three different data sets. i.e. Data sets are trained and tested at base level. It was an effort to find best classifier among these seven classifiers. Also, the accuracies of these base level classifiers will be compared with meta level results.

The results (predictions) of these base level classifiers will also be used as input at meta level in heterogenous ensemble model.

In the next chapters different ensemble techniques are used to classify these data sets. Main goal is to find a model with better accuracy and consistency in accuracy for classification of sentimental text data. For this purpose, homogenous ensemble technique and heterogenous ensemble methods are applied on above mentioned data sets.

Chapter 5 : Homogeneous Ensemble Classifiers

Ensemble classification is considered a better approach and is widely used in classification. Ensemble classification refers to combine different classifiers predictions to classify a new sample. There are two types of ensembles. i.e. Homogenous classifiers and heterogenous classifiers. Homogeneous ensemble uses same type of classifiers at base level. In this chapter, we use homogenous classification methods to classify our data sets. Bagging and boosting are two popular methods used for homogenous ensemble classification.

Homogenous combiners (Ensembles) are used on previously mentioned data sets to find, weather there is significant improvement in the results.

5.1 Boosting

A machine learning ensemble primarily used to minimize bias and variance. Boosting in based on assumption that a combination of week learners (a classifier that is not very highly correlated with true classification i.e. it has a low rate of accuracy) can create a strong learner (a model which is highly correlated with true classification).

There are many boosting algorithms e.g., **AdaBoost, Brown Boost, CoBoost, LpBoost, Gradient Boosting Machine** are to name a few.

The main difference in these boosting algorithms in the method, they weigh the data points and their hypothesis. Most of the algorithms uses the approach to learn week classifiers iteratively with respect to distribution and adding them to find a strong model or a strong final classifier.

GBM (Gradient Boosting Machine) is the one, which is used in this study of find the accuracy on our data sets i.e., IMDB reviews, Yelp reviews and Amazon reviews.

Following figures illustrate how GBM works.

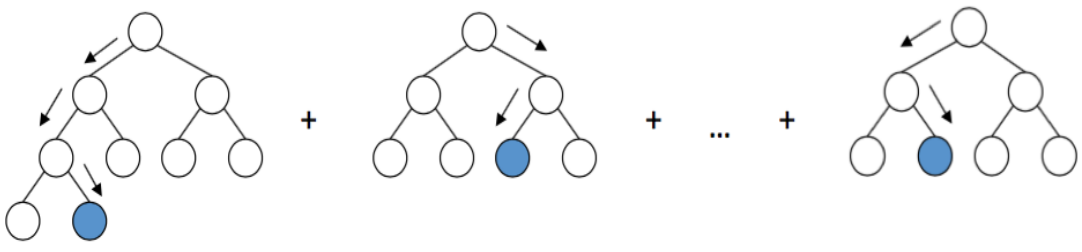

Figure 5.1 Illustrationof how GBM works

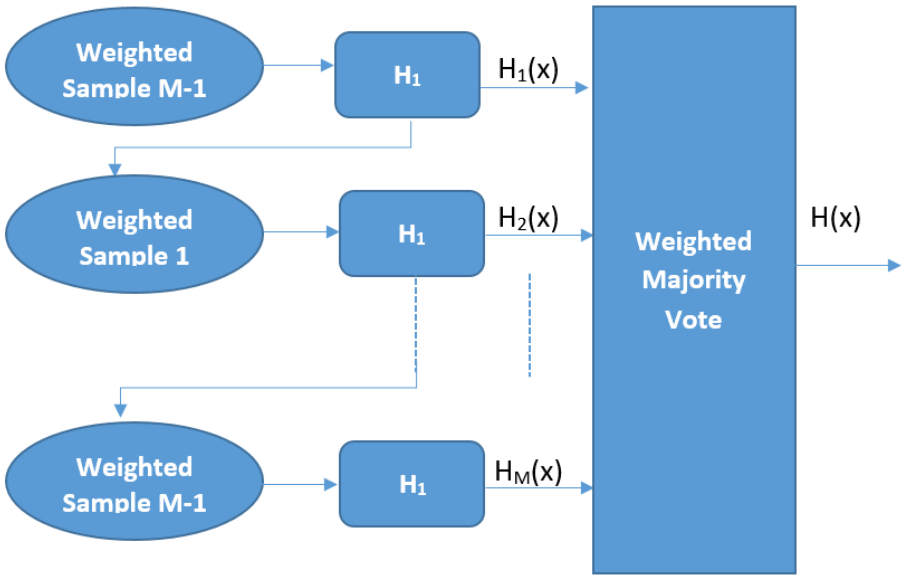

Figure 5.2 GBM's working

5.1.1 Gradient Boosting Algorithm

Input:

- Data sample {(x_1, y)(x_2, y).....x_m, y)}
- Total number of iterations T
- Loss function **L** $(y_i, f(x))$.
- Base learner h (x, Θ).

1 Initialize $\hat{f}_o$ with a constant value f_o (x)=$arg_{min} \sum_{i=1}^{n} L\,(y_i, r)$

2 for i=1 to T

3 Computer negation gradientg_i(x)

4 Apply new base learner h (x, Θ_i)

5 Find best gradient descent step size r_i

6 P_i=$arg_{min} \sum_{j=1}^{n} L\,[y_i, \hat{f}_{i-1}(x_j) + r_h(x_j, \theta_i)]$

7 Update the new estimate of the function $\hat{f}_i = \hat{f}_{i-1} + r_i \mathrm{h}\,(x, \Theta_i)$

8 End for

Computational complexity of GBM in **O (np n_{Tree})** where n in number of training examples P is number of features of training data set and n trees is number of trees generated by this algorithm.

Explanation:

A training data set **D (X_i,Y_j)** where X_i $is\ numbers\ \ of$ data points and Y_j is the class values. In our dataset it is two class problem.

The training loss function is a function that is used to Minimize loss function (which is meaned squared error).

Loss=MSE=$\sum(y_i - y_i^P)^2$

Where y_i**=ith** target value, y_i^P**=ith** prediction & L$\sum(y_i - y_i^P)$is less function. Our objective is to minimize one loss function for our predictions. We continuously fit our base model on over data and analyse data for error.

These errors indicate the data samples which are difficult to fit by the base model. Then we focus on those hard to fit data samples by reducing the errors. In the end we combine all the predictors by adding some weights to each model.

5.2 Bagging

Bagging is ensemble technique to improve stability and accuracy. Bagging is abbreviation of two words, Bootstrap and aggregation. It is combination of two steps.

- Bootstrap sample set.
- Aggregation

Bagging is used to create number of bags of data I.e., we use different subsets of original data set. We select different sample of original dataset randomly with replacement and uniformly. So, any particular data sample can be repeated multiple times. Normally 2/3 rd of training set is selected as random subset. The features or cells are also bagged. Usually $\sqrt{n}$ features are selected randomly of **n** feature space.

Each dataset so selected is then trained over a certain bagging algorithm. Each bag of data is trained on different model. Now we have an ensemble of different model. We query our unknown data sample on each model and use majority voting to find the prediction for required data sample.

The following figure illustrate the bagging process graphically.

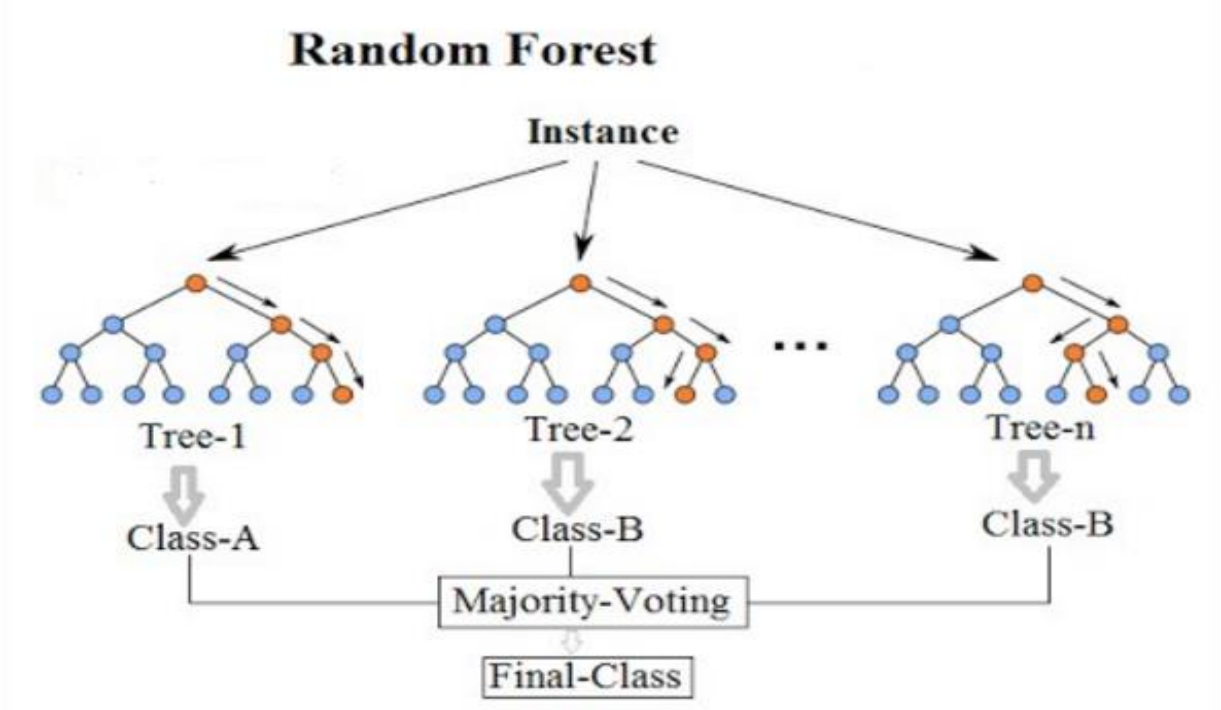

Figure 5.3 Random Forest

Random forest algorithm is used here for bagging purpose to find accuracy on our data sets.

Random Forest was proposed by Breiman in 2001. Algorithm is described below.

Here is brief description, how this algorithm works.

N different data samples are taken from original data set D. These samples can be drawn with or without replacement. We have taken data samples with replacement in this study. So, repetition may take place as these samples are with replacements.

Each data sample is of same size and some instances may repeat more than once while may not appear at all. Each data sample is trained with a model say with weak learner and a classifier is generated. Several classifiers C_i, i=1, 2, 3…n. is generated on n data samples.

Then all classifiers are applied on each sample of test set and majority voting scheme is used to predict the sample's class.

Following is pseudocode in simple steps.

5.2.1 Random Forest Algorithm:

1- Input dataset D

2- Generate n bootstrap data samples D_1, D_2, … D_n

3- Apply Classifier on each data sample.

   $Ci= F_j (D_i)$. F_j is a weak learner and C_i: $X \rightarrow \{0,1\}$: As we have binary class problem.

4- Aggregated classifier of new data instance X' is simply found by taking majority vote

   $C(X') = \text{Majority } (C_i(X'))$ for i=1, 2, …n.

Computational complexity of Random Forest is $O(n^2 p n_{trees})$. Here n is number of training samples and p is number of features, n_{tree} are number of trees generated in the forest.

5.3 Comparison of results with GBM classifier:

5.3.1 GBM Classifier with all datasets of 10000 instances

The figure 5.4 depicts the accuracies of GBM over the selected datasets with 10000 instances. It produces the best results over the yelp dataset which is 89% and its accuracy over IMDB Movie Reviews is 81% whereas the accuracy of this classifier over Amazon Reviews dataset is 80%.

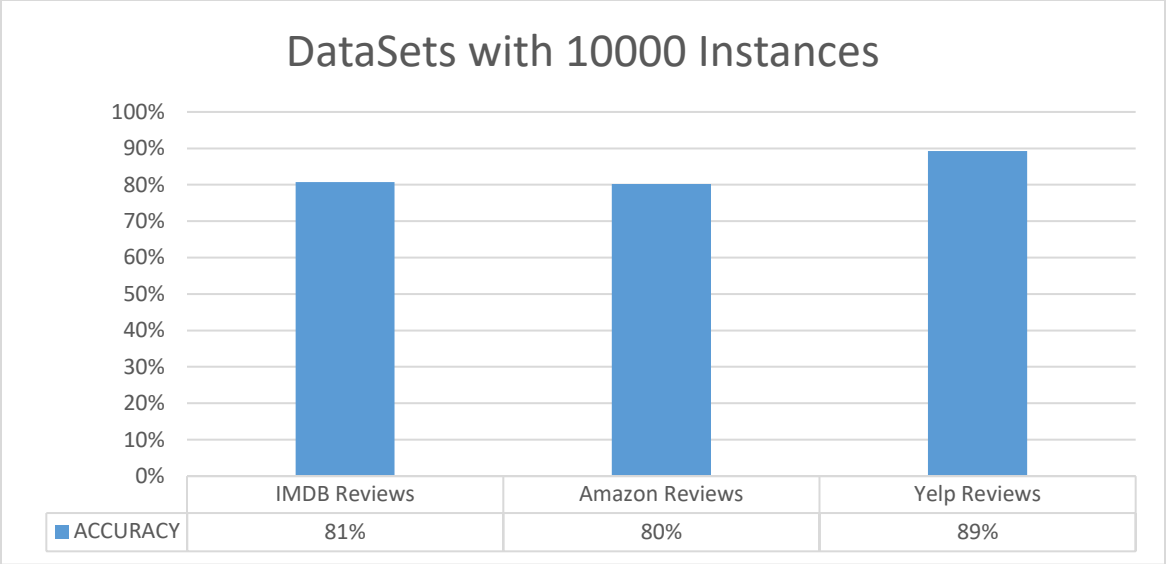

Figure 5.4 Accuracy of GBM classifier on all Datasets with 10000 instances

5.3.2 GBM Classifier with all datasets of 6000 instances

The following figure shows the accuracies of GBM over the selected datasets with 6000 instances. It can be seen that once again it produces the best accuracy over the yelp dataset which is 90% and its accuracy over IMDB Movie Reviews is 81% which same as in the case of 10000 instances whereas the accuracy of this classifier over Amazon Reviews dataset is 78%.

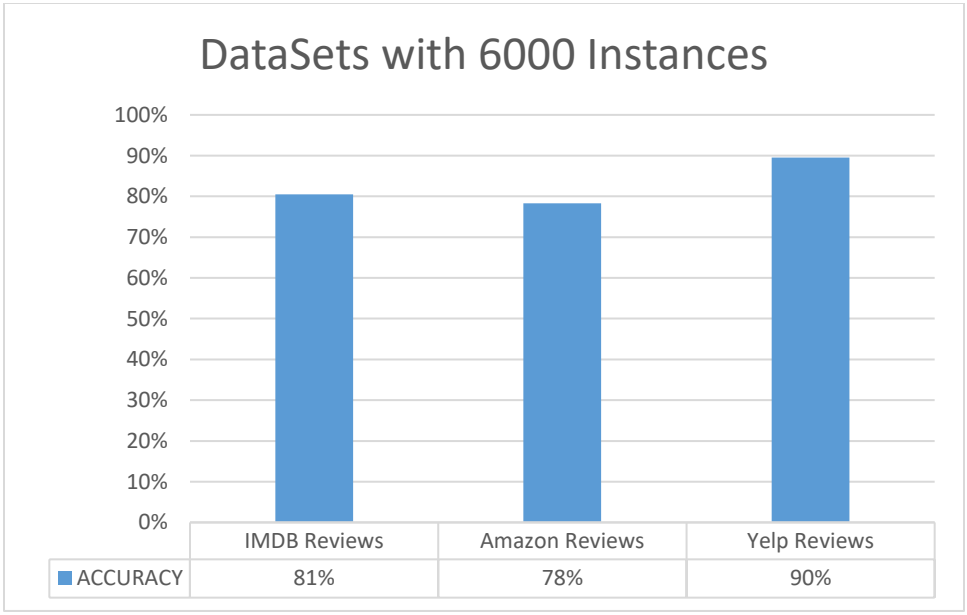

Figure 5.5 Accuracy of GBM classifier on all Datasets with 6000 instances

5.3.3 GBM Classifier with all datasets of 3000 instances

The following figure shows the accuracies of GBM over the selected datasets with 3000 instances. The figure depicts that once again it produces the best accuracy over the yelp dataset which is 91% and improved from both previous datasets. Its accuracy over IMDB Movie Reviews is 79% which is decreased as compared to the previous cases of 10000 and 6000 instances while the accuracy of this classifier over Amazon Reviews dataset is 78%.

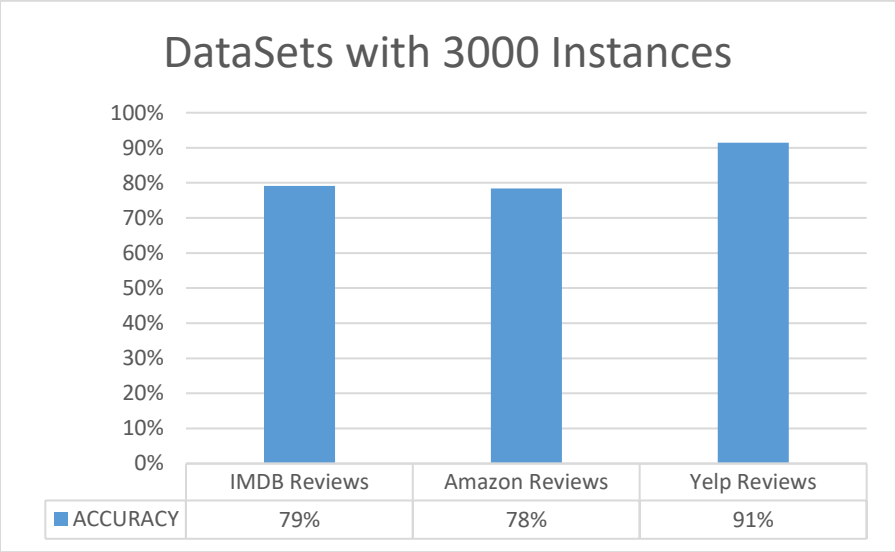

Figure 5.6 Accuracy of GBM classifier on all Datasets with 3000 instances

5.3.4 GBM Classifier with all dataset's of 1500 instances

The following figure shows the accuracies of GBM over the selected datasets with 1500 instances. The figure depicts that once again it produces the best accuracy over the yelp dataset which is 93% and improved from all the previous datasets. Its accuracy over IMDB Movie Reviews is 79% which is decreased as compared to the previous cases of 10000 and 6000 instances but same as in the case of 3000 instances however the accuracy of this classifier over Amazon Reviews dataset is also 79%.

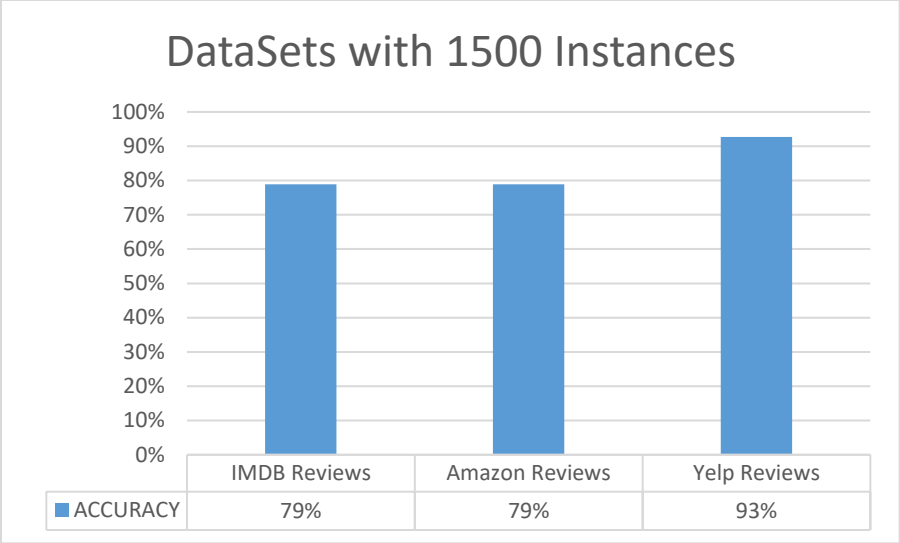

Figure 5.7 Accuracy of GBM classifier on all Datasets with 1500 instances

5.4 Comparison of results with RF classifier:

5.4.1 RF Classifier with all datasets of 10000 instances

The figure 5.8 depicts the accuracies of RF classifier over the selected datasets with 10000 instances. It can be seen that it produces the best results over the yelp dataset which is 86% which around 3% lesser than in the case of GBM with same dataset and its accuracy over Amazon Reviews is 79% whereas it performs worst over the IMDB dataset the accuracy of this classifier over IMDB Reviews dataset is 76%.

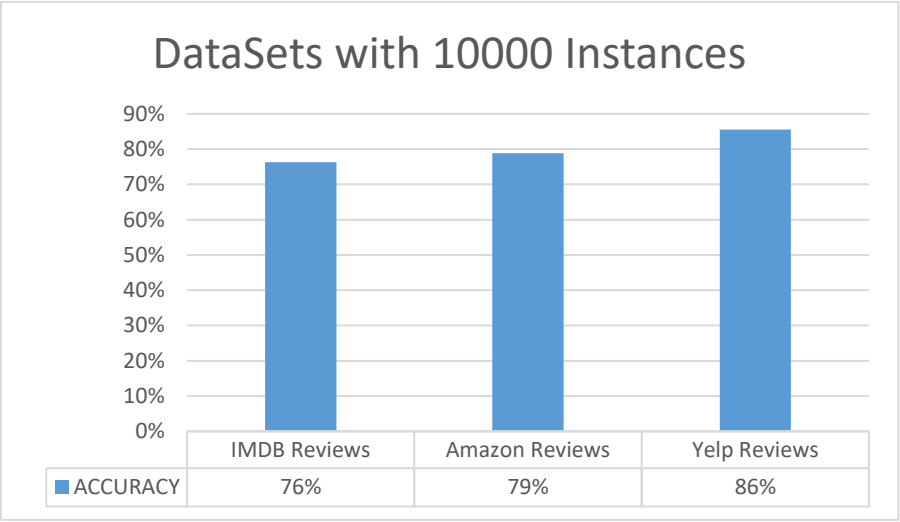

Figure 5.8 Accuracy of RF classifier on all Datasets with 10000 instances

5.4.2 RF Classifier with all datasets of 6000 instances

The following figure shows the accuracies of RF classifier over the selected datasets with 6000 instances. Once again it produces the best accuracy over the yelp dataset which is 88% and its accuracy over IMDB Movie Reviews is 77% which same as in the case of Amazon Reviews dataset.

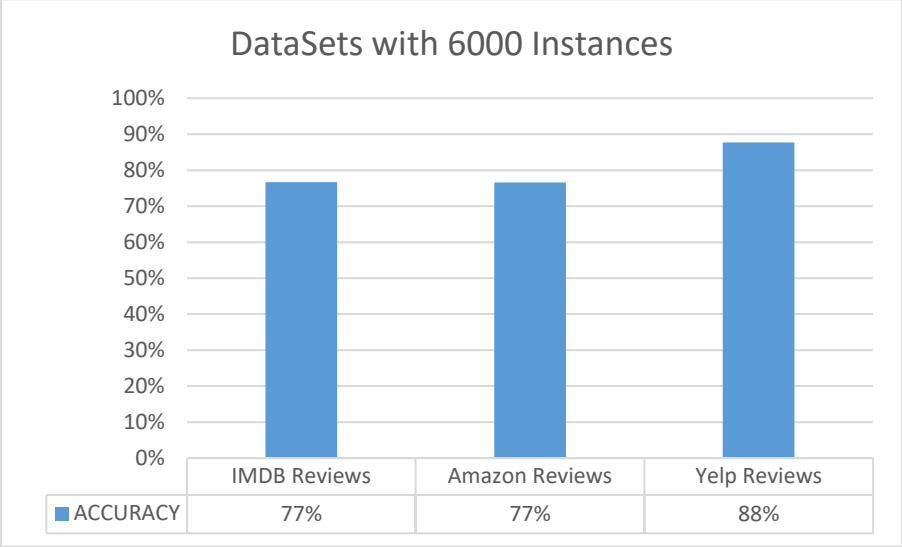

Figure 5.9 Accuracy of RF classifier on all Datasets with 6000 instances

5.4.3 RF Classifier with all datasets of 3000 instances

The following figure shows the accuracies of RF classifier over every dataset with 3000 instances. The figure depicts that once again it produces the best accuracy over the yelp dataset which is 88% although decreased as in the case of GBM but improved from the previous datasets of 10000 instances. Its accuracy over IMDB Movie Reviews is 77% which is same as in the case of 6000 instances and increased compared to the previous case of 10000 instances while the accuracy of this classifier over Amazon Reviews dataset is 78%.

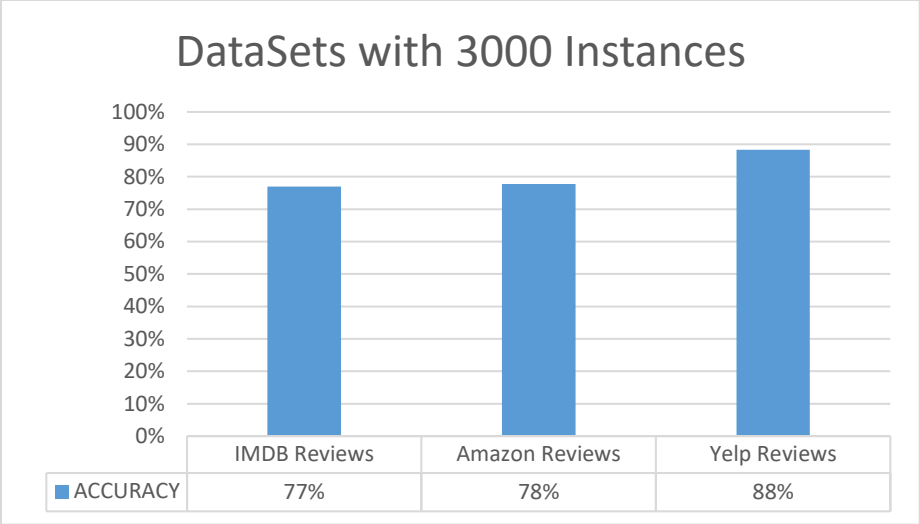

Figure 5.10 Accuracy of RF classifier on all Datasets with 3000 instances

5.4.4 RF Classifier with all data sets of 1500 instances

The following figure shows the accuracies of RF classifier over the selected datasets with 1500 instances. The figure depicts that once again it produces the best accuracy over the yelp dataset which is 90% and improved from all the previous datasets. Its accuracy over IMDB Movie Reviews is 76% which is decreased as compared to the previous cases of 6000 and 3000 instances but same as in the case of 10000 instances however the accuracy of this classifier over Amazon Reviews dataset is also 77%.

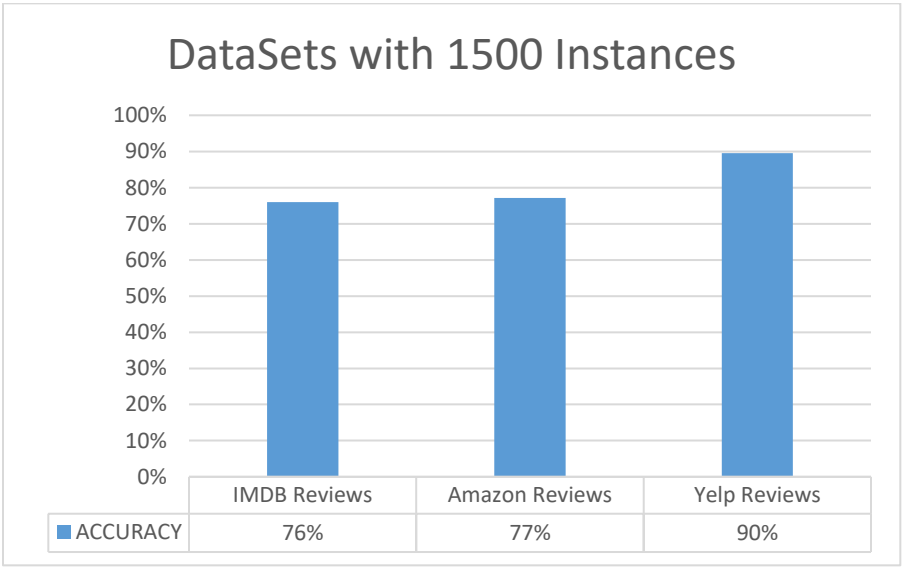

Figure 5.11 Accuracy of RF classifier on all Datasets with 1500 instances

5.5 Comparison of Homogenous classifiers with individual classifiers:

Following are different comparison tables of results of all individual classifiers and homogenous classifiers. Three different measurement tools are used. All the numbers are in percentages. Acc is used for accuracy measure, Sen for sensitivity measure and Spec for specificity measure.

Table 5-1 Accracy Results of homogeneous classifiers on Amazon data set.

Classifier	**10,000 instances**			**6,000 instances**			**3,000 instances**			**1500 instances**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	74	80	68	74	74	73	73	70	76	72	70	73
NB	76	72	80	75	71	79	75	76	74	77	68	85
SVM	80	81	79	78	77	80	**79**	79	78	76	76	76
LDA	80	80	80	78	77	80	78	78	77	77	78	76
RPART	65	95	35	63	29	97	64	95	34	70	87	53
NNET	80	80	80	79	76	81	**79**	78	79	76	74	79
KNN	68	62	74	67	63	71	66	66	81	65	54	76
GBM	**81**	80	80	**81**	76	80	**79**	78	78	**79**	78	80
RF	76	76	81	77	72	81	77	74	82	76	78	76

Table 5-1 shows comparison of results of amazon data set of different sample sizes. Results of individual classifiers and homogenous classifiers are presented in the table. With 10,000 samples GBM, SVM, LDA and NNET have better accuracies than others i.e. 80%. While NNET is performing better with 6,000 samples. SVM and NNET have better accuracy %age with 3,000 samples. And GBM again performed well with 1500 samples. So, no single classifier, whether individual or ensemble performed consistently well on all instances of amazon data set.

Table 5-2 Accuracy Results of homogeneous classifiers on IMDB data set.

Classifier	**10,000 instances**			**6,000 instances**			**3,000 instances**			**1500 instances**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	74	71	78	72	70	73	71	67	76	68	51	84
NB	79	76	82	79	76	81	76	80	72	78	83	73
SVM	**82**	79	84	**81**	80	83	77	76	78	77	78	76
LDA	**82**	78	85	**81**	79	82	**79**	78	80	77	76	78
RPART	67	46	88	68	48	87	67	44	89	65	44	86
NNET	77	77	77	79	79	79	78	71	84	78	77	78

KNN	73	69	79	70	50	88	72	73	70	69	64	73
GBM	80	76	85	78	77	84	78	76	83	**79**	80	78
RF	79	60	93	77	62	92	78	62	92	77	66	86

Results on IMDb data set are presented in table 5-2. It shows that SVM and LDA have better accuracy %age than others on 10,000 samples while GBM outperformed others in all others sample sizes. Again, one cannot figure out which classifier is better performer on all samples of data set.

Table 5-3 Accuracy Results of homogeneous classifiers on YELP data set.

Classifier	**10,000 instances**			**6,000 instances**			**3,000 instances**			**1500 instances**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	84	82	85	81	84	79	81	79	82	82	92	71
NB	81	72	89	81	70	91	51	0.2	99	81	74	88
SVM	**90**	89	91	89	89	89	88	89	88	84	84	85
LDA	89	89	89	89	89	89	89	88	89	85	82	88
RPART	67	38	96	71	92	49	72	93	51	69	40	97
NNET	**90**	90	90	89	89	90	86	85	86	87	86	88
KNN	74	85	62	76	84	68	73	78	68	69	42	96
GBM	89	89	89	**90**	89	90	**91**	90	93	**93**	95	91
RF	86	79	93	88	86	90	88	86	90	90	85	94

Table 5-3 represent accuracy measure on YELP data set. We can see that on a data of 1500 examples of YELP data, GBM outperform all base level classifiers. But on the other hand, LDA and SVM performed better on other data sets of 3,000, 6,000, and 10,000 samples.

Also, GBM has consistency of 67% by getting top position regarding accuracy measure on 8 out of 12 occasions. Clearly GBM is performed better than other base classifiers. So, we can say that boosting is better classification technique for classification of sentiment type data as compared to individual level classifiers.

5.6 Discussion

It can be concluded from the following comparison that by using the GBM and RF (which are the homogeneous ensembles), there is a difference or some improvement in most of cases as compared to the individual classifiers. It can be noticed that GBM outperformed RF in predicting the right class. Although in some particular data sets, homogenous classifiers showed better results and better consistency but it did not perform consistently well on all different samples of all data sets. So, consistency can be a problem while modeling data with homogenous combiners. Therefore, homogenous ensembles cannot be considered as best performing classifiers on binary class sentimental data.

We may achieve far better results by using heterogeneous classifiers. For this purpose stacked generalization model is applied in next chapter in order to obtain better accuracy and consistency.

Chapter 6 : Heterogeneous Ensemble Classifiers

Heterogeneous combiners are applied on same data sets in this chapter i.e. stacking with different classifiers are applied on Amazon, IMDB, YELP data sets and results are compared with earlier mentioned techniques.

6.1 Stacking

Another fascinating method with regards to supervised machine learning is to get familiar with the blend of predictions of base level multiple classifiers. The inspiration driving is to acquire more elevated level of precision at meta level than singular classifier. Stacked generalization or stacking is the method that manages errand of learning meta level classifier to join predictions of various base level classifiers. Accomplishment of stacking is because of assorted classifiers utilized at base level. In this way, decent variety of predictions is accomplished and afterward these predictions are joined. Also, in this manner, a higher exactness can be accomplished at meta level. Likewise, no learning of unique information happens at meta level. Each preparation case in the area of intrigue is spoken to by a vector $<x_n>$ x y, where n is a lot of characteristic qualities or features, and y is the class esteem depicting a specific occasion in the space, which is to be perceived at runtime. So as to order another vector, the predictions of the base-level classifiers structure another feature vector, which is appointed the class an incentive by the meta-level classifier.

The main technique used in this study is stacking (also termed as stacked generalization). The stacking model used in this study is explained below.

In this technique several learning classifiers are trained over given data (training set) and their prediction are combined and a meta data in created then an algorithm (random or of one's choice) is trained on this meta data to make a final prediction. In this study base classifiers **LDA, KNN, RPART, SVM, NB, NNET &CTREE** are used at level 1 to make prediction on Three different data sets i.e., (IMDb reviews, Yelp reviews and amazon reviews). These predictions are combined to form metadata at level-1and is used as input at level 2. At this stage, we used all the classifiers that are used at level 1 iteratively to obtain final prediction. In stacking process, input of stage **2** is the output of stage1. And it is learned by some machine learning algorithm at stage2 and we get a predictor.

Proposed Stacking Model:

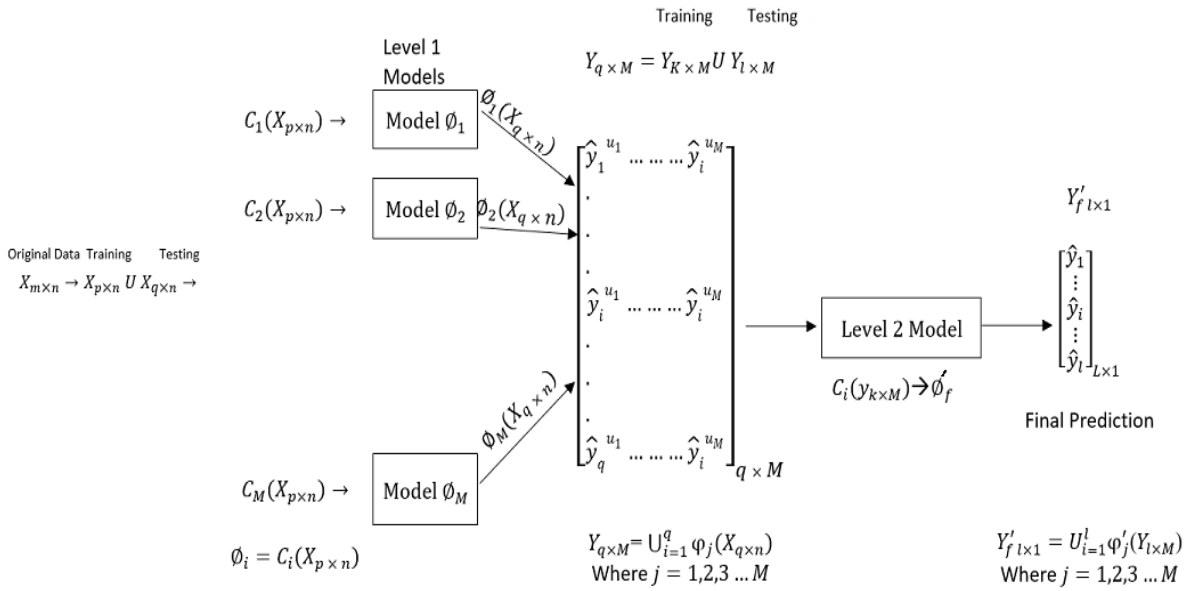

Figure 6.1 Working of Proposed Model

In the figure above, $X_{m\times n}$ represents original Data matrix where **m** is Number of records (row) and **n** is Number of features (columns). Since each word of customer Review is taken as a feature in this data after excluding stop words and applying other pre-conditions:

Let $X_{m\times n}$ be data matrix used for classification purpose. Since data is divided into training and test sets. Let $X_{p\times n}$ be used for training purpose. And $X_{q\times n}$ for test purpose for base level classification. Here $m=p+q$.

Here $X_{m\times n} = X\, U\, Y$, where $X = \cup_{i=1}^{m-1} x_i + Y$, $Y \epsilon \{0,1\}$. In this case $Y = \{0,1\}$. As we have two class problem.

Following model illustrate the process of classification at base level.

$C_i\,(X_{p\times n}) \rightarrow \varphi_i(X_{q\times n}) \rightarrow Y_{q\times M}$. Here $i = 1, 2, \ldots, M$

M are number of classifiers used at base level. $Y_{q\times M}$ is output matrix obtained at base level consist of predictions made by models φ_i on test data.

Now at meta level, data matrix $Y_{q\times M}$ will be used as input. Same base level classifiers are applied at meta level again. So, $Y_{q\times M}$ is again divided inti training and test sets. $Y_{k\times M}$ will serve as training set and $Y_{l\times M}$ will be used as test set. Here $q = l + k$.

$C_i\ (Y_{k\times M}) \rightarrow \varphi_i'(Y_{l\times M}) \rightarrow Y_{f\ l\times 1}'$. Here $i = 1, 2, \ldots, M$

Following is figurative illustration of complete process step by step.

Step-1:

Level 1-Training: M different base models $\phi_1, \phi_2, \ldots, \phi_M$ are trained using dataset $\{X\}_{m\times n}$ and classification algorithms. $C_1, C_2, \ldots, C_M$

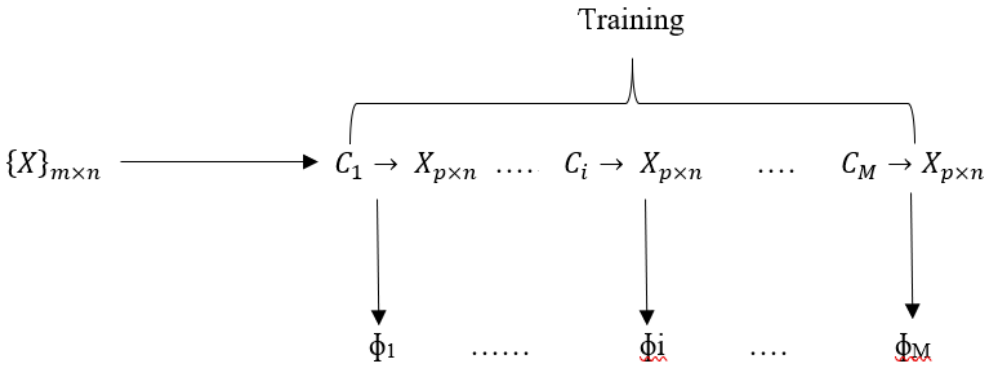

Figure. Base Level Training

Step-2:

Level-1 testing: Trained model $\phi_1, \phi_2 \ldots\ldots \phi_M$ are applied on test dataset $X_{q\times n}$ and M predictions for each data sample is obtained which are then combined to form a new dataset $Y_{q\times M}$.

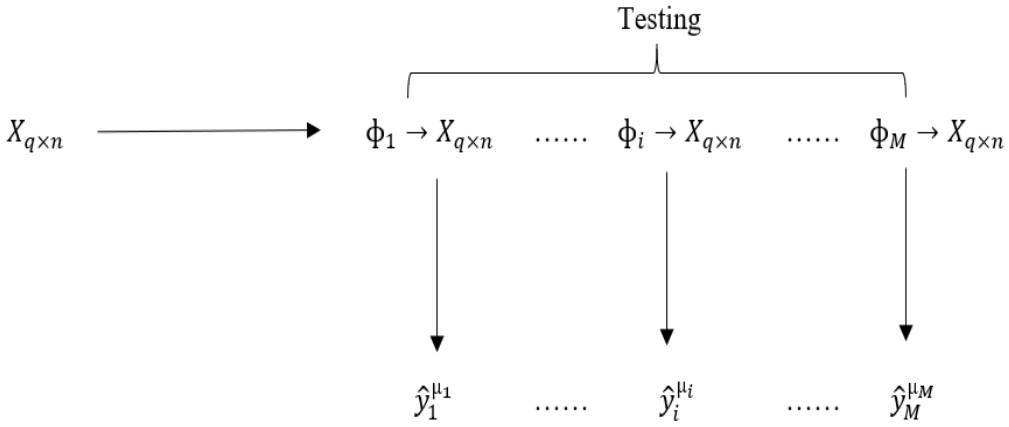

Figure Base level Testing

$$Y_{q\times M} = \begin{bmatrix} \hat{y}_1^{\mu_1} & \cdots & \hat{y}_1^{\mu_M} \\ \vdots & & \vdots \\ \hat{y}_i^{\mu_1} & \cdots & \hat{y}_i^{\mu_M} \\ \vdots & & \vdots \\ \hat{y}_q^{\mu_1} & \cdots & \hat{y}_q^{\mu_M} \end{bmatrix}_{q\times M}$$

i.e. $Y_{q\times M} = \bigcup_{i=1}^{q} \varphi_j(X_{q\times n}) \qquad j = 1,2,3 \ldots M$

Data matrix $Y_{q\times M}$ consist of 0 'S and 1'S only, as it is the prediction of base level classifiers which results in class values. Since datasets we used, are of two class. Where 0 is for negative and 1 is for positive.

Step-3:

Level-2 training: The new formed dataset $Y_{q\times M}$ at **Step 2** is used to train at level-2 i.e.

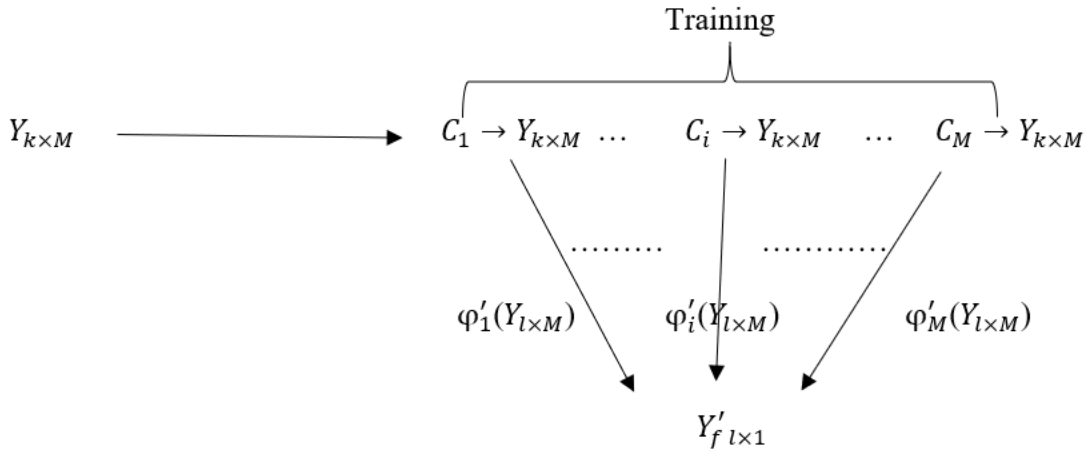

Figure Meta level training

$$Y'_{f\ l\times 1} = U_{i=1}^{l} \varphi'_j(Y_{l\times M}) \qquad j = 1,2,\ldots,M$$

$$Y'_{f\ l\times 1} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_i \\ \vdots \\ \hat{y}_l \end{bmatrix}_{l\times 1}$$

Mathematically whole process can be written as.

Step-1. $C_i(X_{p\times n}) = \varphi_i \qquad i = 1, 2, \ldots, M$

Step-2. $\varphi_i(X_{q\times n}) = Y_{q\times M}$, $Y_{q\times M} = \bigcup_{i=1}^{q} \varphi_j(X_{q\times n})$ is matrix of predictions by base level classification models φ_i on test data.

Step-3 $C_i(Y_{k\times M}) = \varphi'_i \qquad j = 1, 2, \ldots, M$

Step-4 $U_{i=1}^{l} \varphi'_j(Y_{l\times M}) = Y'_{f\ l\times 1}, j = 1, 2, \ldots, M$ Which is final prediction.

6.1.1 Stacking Algorithm:

---

1. //Input training Data $X_{m\times n} = \{Xi, f(Xi)\}$.
2. //Output ensemble classifier H
3. //Step 1: learn base - level classifiers.
4. for i = 1 to M
5. learn ϕ_i on Learning dataset of $X_{m\times n}$
6. end for
7. //Construct new dataset based on prediction of Step 1
8. $X_D = \{\ \}$
9. for i = 1 to m
10. for j = 1 to M
11. $\hat{y} = \phi_j(X_i)$, X_i is Test data
12. End for
13. $X_D = X_D \cup \hat{Y}$
14. End for
15. //Learn a Meta-classifier
16. Learn H= $\phi(X_D)$ a Meta-classifier.
17. Return H

Stacking algorithm have computational complexity of **O(NT$\sum_{i=1}^{n} g_i$ (n))**where g_i **(n)** represent the computational time complexity of base classifier used in the algorithm e.g., if the classifier used is **Naïve Bayes** at meta-level then the running time for whole stacking algorithm will be **O(mTNP).**where N is number of training examples and **P** is number of features of data set.

6.2 Graphical comparison of results with Stacking model:

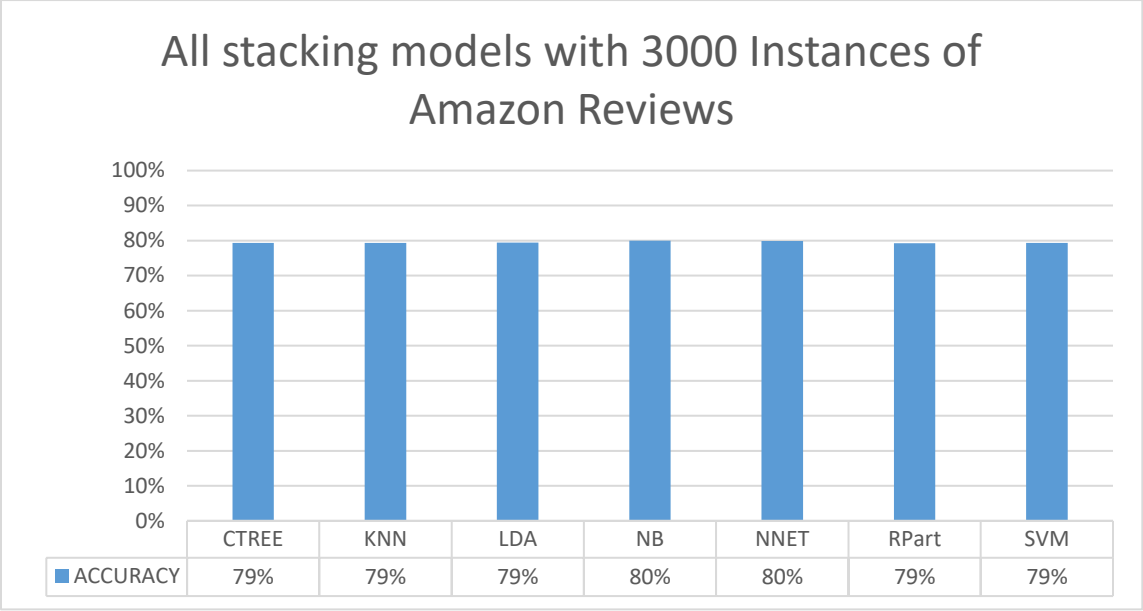

Figure 6.2 Accuracies of all stacked classifiers on Amazon Reviews Dataset with 3000 instances

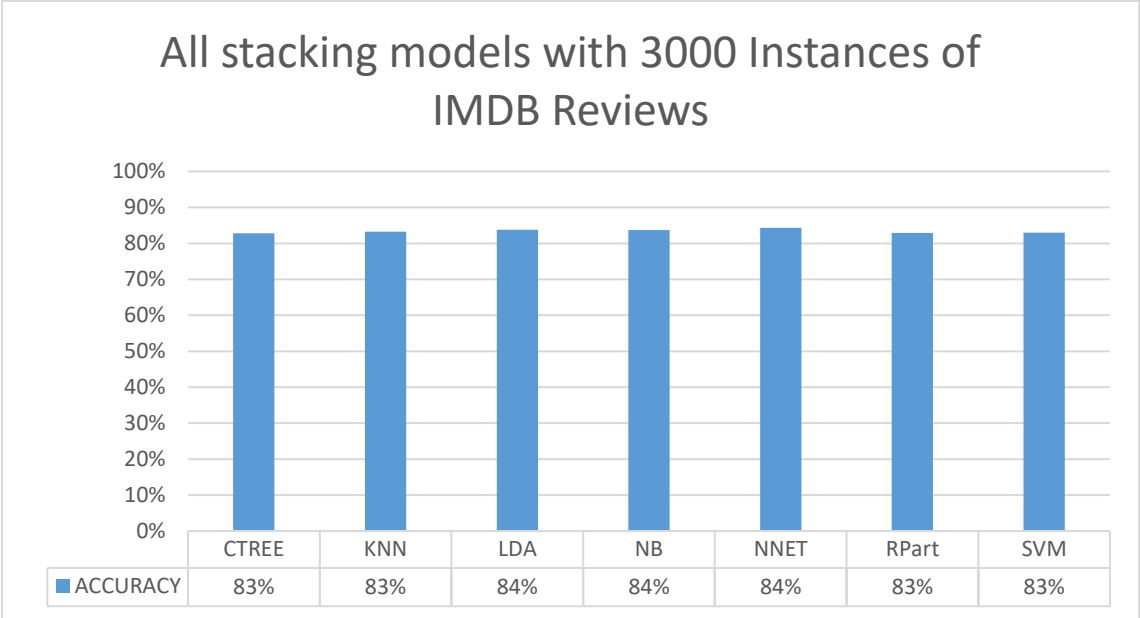

Figure 6.3 Accuracies of all stacked classifiers on IMDB Reviews Dataset with 3000 instances

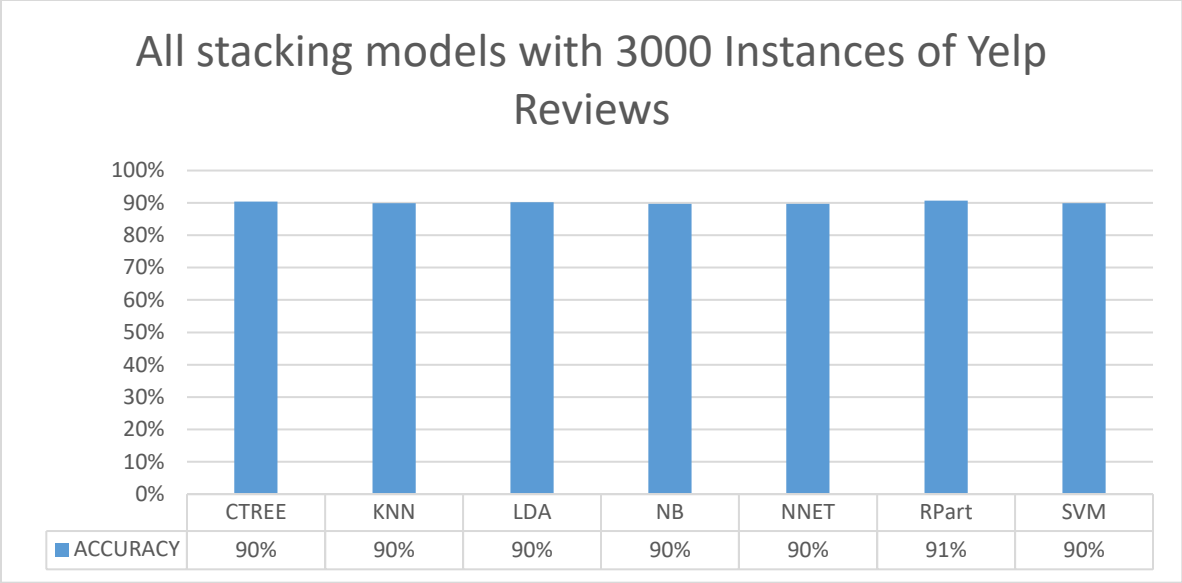

Figure 6.4 Accuracies of all stacked classifiers on Yelp Reviews Dataset with 3000 instances

It can be concluded from the following comparison that by using the GBM and RF which are the homogeneous there is a difference or some improvement in most of cases as compared to the individual classifiers but we may achieve far more better results by using heterogeneous classifiers. Following tables shows the comparison of accuracies of base classifiers and when stacked.

6.3 Comparison tables of base classifiers with stacked model:

Table 6-1 Accuray Results for Amazon of our model on selected data sets of 3000 samples.

Classifier	**Amazon Reviews**			**With Stacking**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	73	70	76	**79**	76	82
NB	75	76	74	**80**	80	80
SVM	79	79	78	**79**	78	81
LDA	78	78	77	**79**	78	80
RPART	64	95	34	**79**	76	82
NNET	79	78	79	**80**	78	82
KNN	66	66	81	**79**	76	83

Table 6-2 Accuray Results for IMDB of our model on selected data sets of 3000 samples.

Classifier	**IMDB Reviews**			**With Stacking**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	74	71	78	**83**	86	82
NB	79	76	82	**84**	80	88
SVM	82	79	84	**83**	78	88
LDA	82	78	85	**84**	80	87
RPART	67	46	88	**83**	75	91
NNET	77	77	77	**84**	79	90
KNN	73	69	79	**83**	79	88

Table 6-3 Accuray Results for YELP of our model on selected data sets of 3000 samples.

Classifier	**YELP Reviews**			**With Stacking**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	84	82	85	**90**	91	90
NB	81	72	89	**90**	89	90
SVM	90	89	91	**90**	89	90
LDA	89	89	89	**90**	90	91
RPART	67	38	96	**91**	91	90
NNET	90	90	90	**90**	90	90
KNN	74	85	62	**90**	91	89

I am discussing accuracy measure in detail in this section. Other measures can also be considered for other reasons. It is clear from above tables that there is significant improvement in accuracy of different datasets. Considering Amazon dataset, we can observe that 13% improvement can be seen when stacked while it was 67% only when individually learned with kNN. Also, improvement can be seen after stacking with all other classifiers when compared with individual classifier. Stacking with SVM and learning with SVM yields same results i.e. 79%. But even then, stacking with NB and NNET provides results with 80% accuracy, which is improvement of 1% then best individual classifier's accuracy i.e. 79% by SVM and NNET.

Now take second dataset IMDB. It is clear from above tables that there is quite a significant improvement in accuracy measure. Comparing with individual classifiers, we see a lot of improvement. The maximum accuracy achieved by any individual classifier is 79%. On the other hand, max accuracy achieved with stacking is 84%, so minimum of 5% improvement in accuracy is achieved with IMDB dataset. While max of 16% difference in accuracies can be observed, when we use RPART classifier individually and while stacking with RPART. And minimum of 5% improvement can be seen while using LDA individually and while stacking with LDA.

Now let's consider third dataset YELP. Here min improvement is 1% when LDA is used individually (89%) and after stacking with LDA (90%). And maximum improvement is 19% while stacking with RPART (91%) and individually (72%). Also, minimum improvement is 2% Considering overall scenario, i.e. Max improvement obtained by individual classifier is 89% (LDA). And maximum accuracy achieved with stacking is 91% (with RPART).

Following table compares the accuracies of base level classifiers with 1500 samples and accuracy with 1800 samples with stacking.

Table 6-4 Accuray Results for Amazon of our model on selected data sets of 1800 samples. Accuracy measures of base level classifiers are with 1500 samples.

Classifier	**Amazon Reviews 1500 instances**			**With Stacking**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	72%	70	73	**78%**	75	81
NB	77%	68	85	**77%**	72	82
SVM	76%	76	76	**77%**	78	81
LDA	77%	78	76	**79%**	75	82
RPART	70%	87	53	**77%**	73	80
NNET	76%	74	79	**77%**	72	82
KNN	65%	54	76	**79%**	76	81

Table 6-5 Accuray Results for IMDB of our model on selected data sets of 1800 samples. Accuracy measures of base level classifiers are with 1500 samples.

Classifier	**IMDb Reviews 1500 instances**	**With Stacking**

	Acc	Sen	Spec	Acc	Sen	Spec
CTREE	68%	51	84	**82%**	79	86
NB	78%	83	73	**81%**	76	87
SVM	77%	78	76	**81%**	78	88
LDA	77%	76	78	**81%**	80	83
RPART	65%	44	86	**81%**	75	88
NNET	78%	77	78	**81%**	79	83
KNN	69%	64	73	**82%**	80	85

Table 6-6 Accuray Results of for YELP our model on selected data sets of 1800 samples. Accuracy measures of base level classifiers are with 1500 samples.

Classifier	**YELP Reviews 1500 instances**			**With Stacking**		
	Acc	**Sen**	**Spec**	**Acc**	**Sen**	**Spec**
CTREE	82%	92	71	**89%**	93	86
NB	81%	74	88	**90%**	93	88
SVM	84%	84	85	**90%**	89	90
LDA	85%	82	88	**90%**	91	89
RPART	69%	40	97	**93%**	93	86
NNET	87%	86	88	**91%**	91	90
KNN	69%	42	96	**90%**	90	90

Considerable improvement in accuracy can be observed at maximum instances. An improvement of 24% is seen when RPART is stacked. Also, improvement in accuracy is consistent for different for different instance sizes.

6.4 Comparison of Base level classifier with Stacked classifier using AUC-ROC curve

As already mentioned, ROC-AUC is another reliable measure for measuring performance of classifiers. Following are ROC curves and AUC of two different classifiers. Figure 1 shows AUC-ROC of NB classifier on Amazon data set of 3,000 examples at base level. While figure 2 shows AUC-ROC of NB on stacked data.

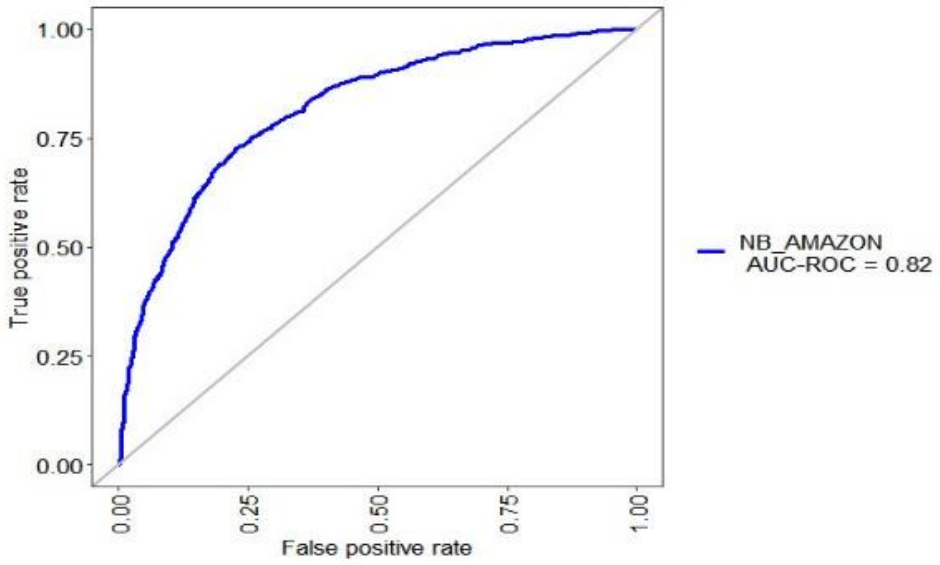

Figure 6.5 ROC for Naïve Bayes (individual) model

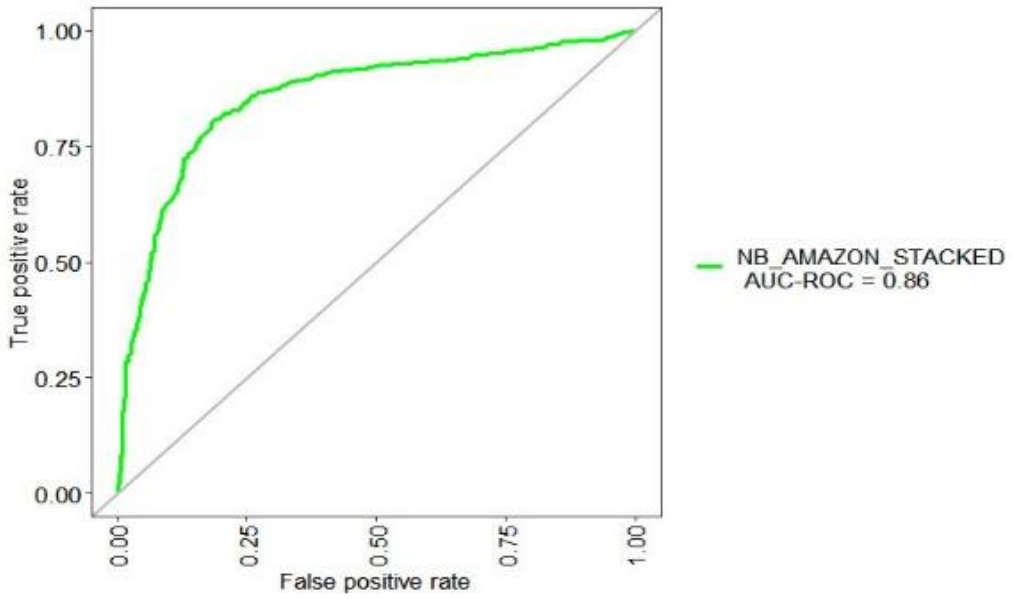

Figure 6.6 ROC for Naïve Bayes (stacked) model

It is clear from above figures that AUC for NB is 0.82 at base level. And 0.86 after stacking which is much improved. Following figure shows comparison of NB at base level and at stacking level.

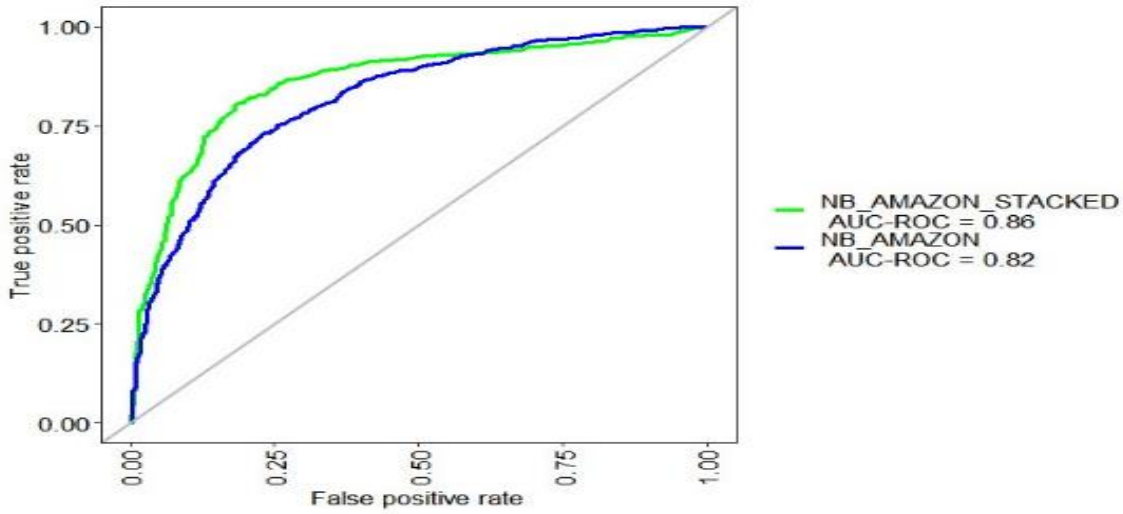

Figure 6.7 Comparison between ROC for Naïve Bayes (individual and stacked) models

Above figure confirms the results mentioned in accuracy tables. It can be seen that probability of getting accurate prediction is better after model stacking than at base level.

Following are AUC-ROC curves for knn classifier at base level and after stacking. It also shows a clear improvement in accuracy and probability of finding correct prediction of a sample query after stacking than an individual classifier.

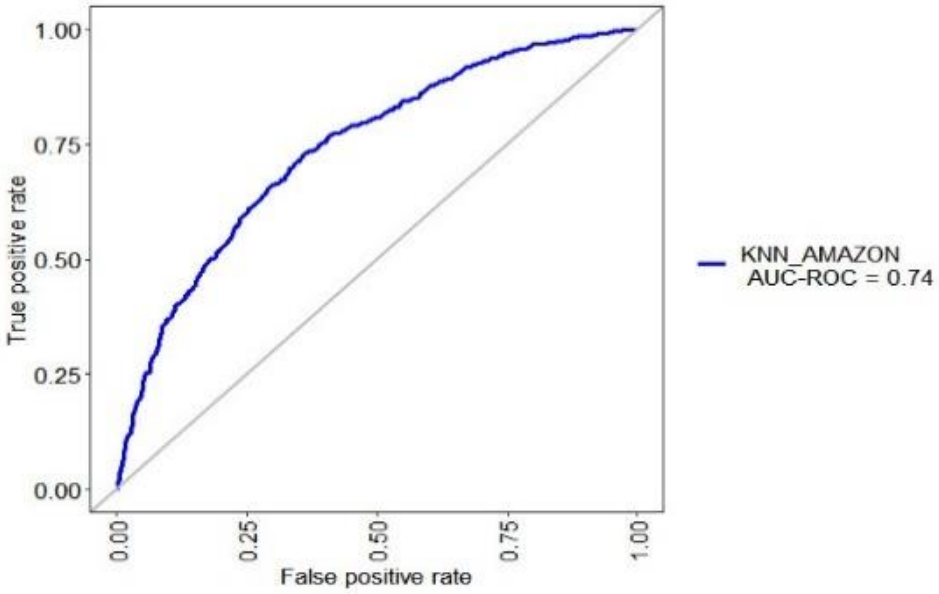

Figure 6.8 ROC for knn (individual) model

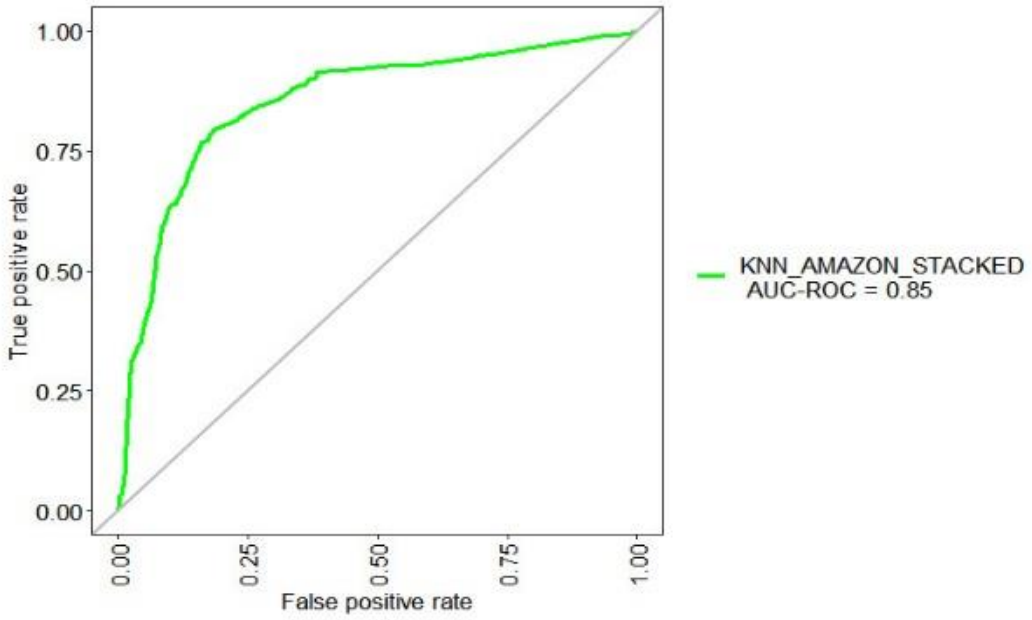

Figure 6.9 ROC for knn (stacked) model

Following figure also confirms the results shown previously in accuracy tables. This result can be extended to all classifiers used at base level.

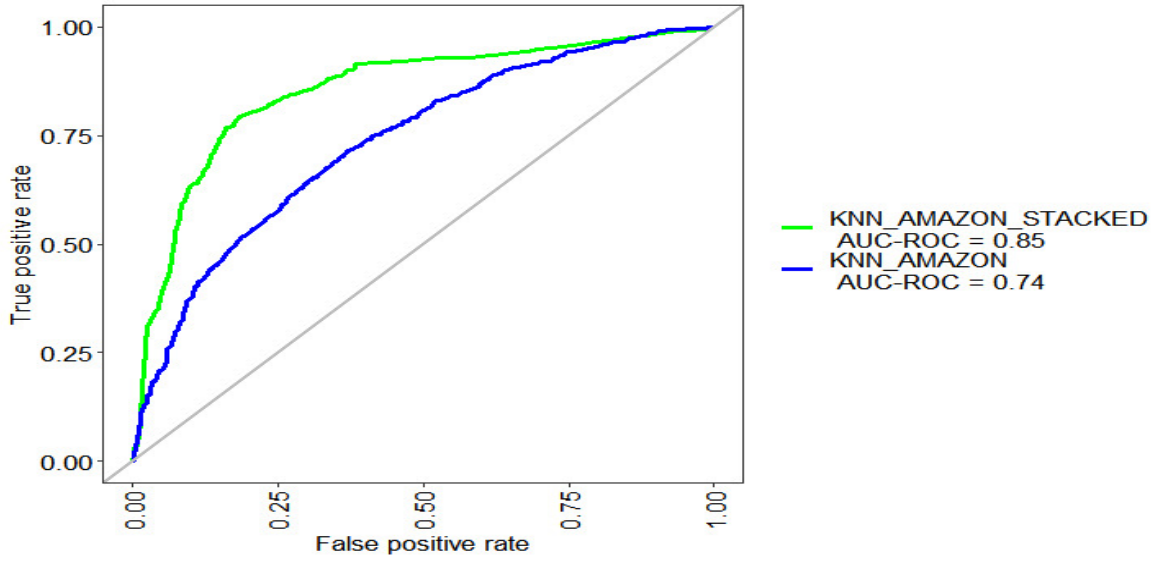

Figure 6.10 ROC for knn (individual and stacked) models

6.5 Discussion

It is clear from above comparison tables that better results can be achieved using stacked generalization. Also, the best accuracy achieved with any of sample size can be improved or at least retained when stacked with any of classifier at meta level. With stacking both the objectives i.e. improvement in accuracy and consistency can be achieved.

Since, different sample sizes of three different datasets have been used. So, we can generalize our conclusion to all sentimental type data.

Following steps must be followed to implement the model verified in this study for improved accuracy results for sentimental data consisting of two classes.

- Different heterogenous base level classifiers must be taken. i.e. classifiers from different family must be selected at base level to learn sentimental data.
- Outcome of these base level classifiers must be combined to form a new data set, which will be input at next level.
- Any of base level classifier can be used at meta level for learning this input data to form a model. It is recommended that the classifier with best accuracy should be used at meta level to obtain best results.

Chapter 7 : Conclusions and Future Work

Text data from micro-blogs and forums in addition to news source, is broadly used in Sentiment Analysis. The subjected information is of great importance in articulating and expressing public's feelings and opinions related to a certain topic or product. But data from social network and micro-blogging websites requires deeper investigation and analysis. Recently researchers have been using Natural Language Processing tools to strengthen the process of Sentiment Analysis, but it still needs some enhancements.

7.1 Conclusion

The main purpose of this research work was to address accuracy improvement and consistency in accuracy on different sentiment data sets. There are certain models which shows excellent results on one dataset but do not perform that well on another dataset. So, consistency in accuracy improvement has been a challenge for data scientists. This research work presents the experimentation containing diversified pre-processing techniques which provide an effect on the presentation of classification of the IMDB movie reviews, Amazon Reviews, Yelp dataset. All the datasets were pre-processed by different techniques and the effects of these pre-processing steps on the performance of classification algorithm in term of accuracy were examined. Our results show that by elimination of the meaningless words, stop words, numbers, and word less than 3 alphabets, satisfactorily increased the enactment of the classification algorithms as, it reduces the feature set of data matrix such formed and thus reducing the computational time complexity.

The technique used, increases the expressiveness of algorithms which can be used to make them further appropriate for stacking. The experiential evaluation of our methodology presented that it leaves behind conventional algorithms in terms of accuracy as well as is much more concise. I look this as a footstep towards an understandable idea of combining multiple models. In comparison of most of the existing work uses non-symbolic learning methods (e.g. neural networks) where were numerous apparent directions for further work.

Same is the case with base level classifiers, as some of them perform better on some instances of data while other perform better on different dataset. Experimental results also show that proposed methodology in this research out perform all other methods used in this research.

An exclusive heterogeneous ensemble model for classification of sentiment data is presented in this research. The investigational results demonstrate that the projected solution is specifically

effective and efficient. It represents its capability for the strategic combination of different classification algorithms. The model is also computationally efficient. Though, Computational time complexity matter can an area of concern if there are an increasing number of base classifiers to be enclosed in the ensemble together with large dataset. The choice of the basic classifier considers the individual learners and the complexity of each of the inference algorithms and leads to a reasonable trade-off between the contribution in terms of accuracy and the associated computation time.

Sentiment analysis has emerged as an interesting area with many problems because it includes natural language processing. There are a range of applications which can benefit from the results, such as news analysis, marketing analysis, tracking customer, competitor analysis and answering questions. Gaining significant intuitions from opinions communicated on the Internet, especially from social media blogs, is essential for many companies and institutions, including product feedback, public opinion, and investor opinion. I also examined movie review sentiment. I used an amalgamation of various pre-processing approaches to decrease text noise and remove extraneous features that don't affect orientation. Extensive experimental results have been shown, demonstrating that the accuracy of proper text preprocessing achieved with the selected datasets is comparable to the accuracy achievable with topic classification and is a much simpler problem.

7.2 Limitations

The datasets used in this research are based on reviews of customers. Data bases of Amazon, YELP and IMDB are used for this purpose. Since reviews used are of limited length so, the proposed model may not be suitable for larger documents. Lager length of document may result in a greater number of attributes of data frames. Which may increase computational complexity of the mentioned model. Also, it can affect accuracy of classification.

Vocabulary constraints are another area which is not addressed in this research. The words “happy” and “not happy” may be treated in same meaning i.e. “happy”. As, not may be discarded as a stop word during preprocessing steps.

7.3 Future work

Future work will explore other sequence learning models for target expression detection and further evaluate the approaches proposed in other languages and domains. This work has opened many future research venues. For example, you can draw on this research by applying and testing the recommended framework with other sentiment analysis techniques. What's more, the element choice can be stretched out to incorporate other n-gram highlights, for example, 2 and 3 grams. It

is intriguing to consolidate such highlights with semantic implications by applying semantic parsers in future investigations.

Finding correlation between number of attributes and number of records can be another field of interest for researchers. As it is seen in this study that reducing number of attributes improves accuracy.

To analyze a sentimental expression or to convert it into tokens is a tricky job. As it can be quite a misleading. And it can also affect the accuracy. For example, when a word “happy” and “not happy” is converted into category with “not” as stop word, it becomes same attribute in sample and can affect the predictive value of class .So, it will be another area to explore for researchers that how these kinds of problems should be handled in natural language processing. Next our study comprises of two class problem. It can be extended to multi class data of more than two classes.

References

[1] E. K. S. A. D. e. a. Kim, "BMC Bioinformatics," 14 February 2009. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-260?optIn=false.

[2] G. &. E. J. F. Seni, Ensemble methods in data mining: improving accuracy through combining predictions., Synthesis lectures on data mining and knowledge discovery, 2010.

[3] R. T. Clemen, "Combining forecasts: A review and annotated bibliography," *International journal of forecasting,* vol. 5, no. 4, pp. 559-583, 1989.

[4] M. Kantardzic, Data mining: concepts, models, methods, and algorithms., IEEE, John Wiley & Sons., 2011.

[5] B. Liu, "Sentiment Analysis and Opinion Mining," in *Sentiment Analysis and Opinion Mining*, 2012.

[6] B. Liu, Handbook of Natural Language Processing, 2010.

[7] M. B. J. T. M. V. K. &. S. M. Taboada, "Lexicon-based methods for sentiment analysis," *Computational linguistics,* vol. 37, no. 2, pp. 267-307, 2011.

[8] K. Z. &. M. N. N. Aung, "Sentiment analysis of students' comment using lexicon based approach," in *IEEE/ACIS 16th international conference on computer and information science (ICIS)*, 2017.

[9] K. T. K. V. L. S. T. &. G.-T. A. Van Nguyena, "Enhancing lexical-based approach with external knowledge for Vietnamese multiple-choice reading comprehension," *arXiv:2001.05687,* 2020.

[10] X. L. A. a. G. A. Zhang, "A Lightweight Online Advertising Classification System using Lexical-based Features," in *Proceedings of the 14th International Joint Conference on e-Business and Telecommunications (ICETE 2017)*, 2017.

[11] K. T. L. J. P. N. K. S. Hong J., "Phishing URL Detection with Lexical Features and Blacklisted Domains," *Adaptive Autonomous Secure Cyber Systems,* 2020.

[12] A. P. M. S. a. V. V. Aditya Joshi, "Towards Sub-Word Level Compositions for Sentiment Analysis of Hindi-English Code Mixed Text," in *26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, 2016.

[13] V. E. Fermi, "Sentiment Analysis in Social Media Texts," in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity*, Atlanta, Georgia, 2013.

[14] F. H. H. H. H. K. L.-B. Chris Thornton, "Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms," in *KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, Chicago, Illinois, USA, 2013.

[15] T. G. &. M. Krzyśko, "Regression Methods for Combining Multiple Classifiers," *Communications in Statistics – Simulation and Computation,* vol. 44, pp. 739-755, 2015.

[16] J. B. G. L. Liu, "Peer-Dependence Valuation Model for Real Estate Appraisal," *Data-Enabled Discovery and Applications,* 2019.

[17] F. Y.-V. MeteOzaya, "Hierarchical distance learning by stacking nearest neighbor classifiers," *Information Fusion,* vol. 29, pp. 14-31, 2016.

[18] A. B. M. a. O. F. Danesh, "Improve text classification accuracy based on classifier fusion methods," in *Information Fusion, 2007 10th International Conference on. IEEE*, 2007.

[19] C. P. M.Sivakumar, "A Hybrid Text Classification Approach Using KNN And SVM," *International Journal of Innovative Research in Science, Engineering and Technology,* vol. 3, no. 3, pp. 1987-1991, 2014.

[20] S.-S. S.-F. L.-C. J. S. J.-F. L.E.Agustı´n-Blas, "A new grouping genetic algorithm for clustering problems," *Expert Systems with Applications,* vol. 39, no. 10, pp. 9695-9703, 2012.

[21] R. Sikora, "A modified stacking ensemble machine learning algorithm using genetic algorithms," in *Handbook of Research on Organizational Transformations through Big Data Analytics*, IGI Global, 2015, pp. 43-53.

[22] R. S. R. S. S. a. H. E. Sexton, "Improving decision effectiveness of artificial neural networks: a modified genetic algorithm approach," *Decision Sciences,* vol. 34, no. 3, pp. 421-442, 2003.

[23] B. B. I. G. Nikolaos Mastrogiannis, "A method for improving the accuracy of data mining classification algorithms," *Computers & Operations Research,* vol. 36, pp. 2829-2839, 2009.

[24] G. T. Harshita Patel, "A Hybrid Weighted Nearest Neighbor Approach to Mine Imbalanced Data".

[25] S. Y. Rita Mandal, "An Improved Intrusion System Design using Hybrid Classification Technique".

[26] G. P. M. W. L. K. Shengyi Jiang, "An improved K-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications,* vol. 39, p. 1503–1509, 2012.

[27] E. T. Huy Nguyen Anh Pham, "A meta-heuristic approach for improving the accuracy in some classification algorithms," *Computers & Operations Research,* p. 174–189, 2011.

[28] Y. H. &. J. A. K. Li, "Classification of text documents," *The Computer Journal,* vol. 41, no. 8, pp. 537-546., 1998.

[29] D. L. M. A. (. A. N. A. a. T. S. I. C. S. M. S. M. C. S. P. P. S. Haque S.M., "A News Analysis and Tracking System," *Pattern Recognition and Machine Intelligence,* vol. 5909, 2009.

[30] I. K. J. C. K. V. P. G. &. S. C. D. Androutsopoulos, "An evaluation of naive bayesian anti-spam filtering," in *11th European Conference on Machine Learning*, Barcelona Spain, 2000.

[31] L. S. Larkey, "Automatic essay grading using text categorization techniques," in *21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998.

[32] B. &. L. L. Pang, "Opinion mining and sentiment analysis. Foundations and Trends," *Information Retrieval,* vol. 21, no. 1-2, pp. 1-135, 2008.

[33] M. S. M. A. M. N. M. I. T. K. D. D. K. I. &. M. Z. Hossain, "Can ensemble techniques improve coral reef habitat classification accuracy using multispectral data?.," *Geocarto International,* pp. 1-19, 2019.

[34] R. T. H. J. F. a. H. M. W. Joshua S. White, "Hybrid Sentiment Analysis Utilizing Multiple Indicators To Determine Temporal Shifts of Opinion in OSNs," *Cyber Sensing,* 2016.

[35] X.-J. Z. Sharareh R. Niakan Kalhori, "Improvement the Accuracy of Six Applied Classification Algorithms through Integrated Supervised and Unsupervised Learning Approach," *Journal of Computer and Communications,* vol. 2, pp. 201-209, 2014.

[36] A. Verma and S. Mehta, "A comparative study of ensemble learning methods for classification in bioinformatics," in *7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, Noida, 2017.

[37] F. L. C. F. J. O. C. G. V. J. A. T. Fernando Enríquez, "A comparative study of classifier combination applied to NLP tasks," *Information Fusion,* vol. 14, pp. 255-267, 2013.

[38] R. a. M. E. Mousavi, "A new ensemble learning methodology based on hybridization of classifier ensemble selection approaches," *Applied Soft Computing,* vol. 27, pp. 652-666, 2015.

[39] H. S. a. M. K. Fatemeh Nemati Koutanaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *Journal of Retailing and Consumer Services,* vol. 27, no. C, pp. 11-23, 2015.

[40] P. Shrivastava and M. Shukla, "Comparative analysis of bagging, stacking and random subspace algorithms," in *International Conference on Green Computing and Internet of Things (ICGCIoT)*, 2015.

[41] M. E. R. Mousavi, "A new ensemble learning methodology based on hybridizationof classifier ensemble selection approaches," *Applied Soft Computing,* vol. 37, p. 652–666, 2015.

[42] A. S. M. S. R. I. H. F. Thabit Sabbah, "Hybridized term-weighting method for Dark Web classification," *Neurocomputing,* 2015.

[43] J. S. J. M. K. X. J. G. Gang Wang, "Sentiment classification: The contribution of ensemble learning," *Decision Support Systems,* 2013.

[44] E. S. Lee, "Exploring the Performance of Stacking Classifier to Predict Depression Among the Elderly," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, Park City, UT, USA, 2017.

[45] A. R. T. Anjali Gupta, "Optimization of stacking ensemble Configuration based on various metahueristic algorithms," in *IEEE International Advance Computing Conference*, 2014.

[46] C. J. Merz, "Using correspondence analysis to combine classifiers," *Machine Learning,* vol. 36, pp. 33-58, 1999.

[47] D. H. Wolpert, "Stacked generalization," *Neural networks,* vol. 5, no. 2, pp. 241-259, 1992.

[48] D. Zhu, "A hybrid approach for efficient ensembles," *Decision Support Systems,* vol. 48, no. 3, pp. 480-487,

2010.

[49] G. P. C. D. S. M. H. Georgios Sigletos, "Combining Information Extraction Systems Using Voting and Stacked Generalization," *Journal of Machine Learning Research,* vol. 6, pp. 1751-1782, 2005.

[50] A. a. D. J. Fast, "Why stacked models perform effective collective classification," in *Data Mining ICDM'08. Eighth IEEE International Conference on*, 2008.

[51] P. N. B. T. D. Horvitz, "The Combination of Text Classifiers Using Reliability Indicators," *Information Retrieval,* vol. 8, no. 1, p. 67–100, 2005.

[52] L. Y. EitanMenahem, "Troika – An improved stacking schema for classification tasks," *Information Sciences,* vol. 179, no. 24, pp. 4097-4122, 2009.

[53] J. Chen, C. Wang and R. Wang, "Using Stacked Generalization to Combine SVMs in Magnitude and Shape Feature Spaces for Classification of Hyperspectral Data," *IEEE Transactions on Geoscience and Remote Sensing* , vol. 47, no. 7, pp. 2193 - 2205, 2009.

[54] S. Li, C. Zong and X. Wang, "Sentiment Classification through Combining Classifiers with Multiple Feature Sets," in *International Conference on Natural Language Processing and Knowledge Engineering*, 2007, 2007.

[55] J. Ireneusz Czarnowski, "Stacking-Based Integrated Machine Learning with Data Reduction," in *International Conference on Intelligent Decision Technologies*, 2017.

[56] I. Czarnowski and P. Jedrzejowicz, "Stacking and rotation-based technique for machine learning classification with data reduction," in *International Conference on INnovations in Intelligent SysTems and Applications*, 2017.

[57] I. J. P. Czarnowski, "Learning from examples with data reduction and stacked generalization," *Journal of Intelligent & Fuzzy Systems,* vol. 32, no. 2, pp. 1401-1411, 2017.

[58] C. Z. S. L. &. X. G. Yuan Tian, "Multiple Classifier Combination For Recognition Of Wheat Leaf Diseases," *Intelligent Automation and Soft Computing,* vol. 17, no. 5, pp. 519-529, 2011.

[59] "Hybrid classification algorithms for terrorism prediction in Middle East and North Africa," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),* vol. 4, no. 3, pp. 2278-6856, 2015.

[60] M. B. B. G. Christiane Lemke, "Metalearning: a survey of trends and technologies," *Artif Intell Rev,* vol. 44, p. 117–130, 2015.

[61] "Why Stacked Models Perform Effective Collective Classification," in *Eighth IEEE International Conference on Data Mining*, 2008.

[62] S. J. S. G. V. &. D. D. Tulyakov, "Review of classifier combination methods," *Machine learning in document analysis and recognition,* vol. 90, pp. 361-386, 2008.

[63] S. K. Thompson, Sampling, 2012.

[64] R. M. I. W. O. R. a. G. K. Scheaffer, Elementary survey sampling, 2011.

[65] K. Xu, "Mining comparative opinions from customer reviews for Competitive Intelligence.," *Decision support systems,* vol. 50, no. 4, 2011.

[66] Q. L. Q. &. D. R. A. Miao, "A sentiment mining and retrieval system," *Expert Systems with Applications,* 2009.

[67] M. S. a. Y.-F. W. Elli, "Amazon Reviews," *business analytics with sentiment analysis,* 2016.

[68] R. Parveen, N. Shrivastava and P. Tripathi, "Sentiment Classification of Movie Reviews by Supervised Machine Learning Approaches Using Ensemble Learning & Voted Algorithm," in *2nd International Conference on Data, Engineering and Applications (IDEA)2020*, 2020.

[69] S. K. g. ,. H. B. Aytug Onan, "A Multiobjective Weighted Voting Ensemble Classifier Based on Differential Evolution Algorithm for Text Sentiment Classification," *Expert Systems with Applications,* 2016.

[70] T. a. D. D. Shaikh, "Feature Selection Methods in Sentiment Analysis and Sentiment Classification of Amazon Product Reviews," *International Journal of Computer Trends and Technology (IJCTT),* 2016.

[71] J. L. K. T. M. B. R. A. &. S. N. Leevy, "A survey on addressing high-class imbalance in big data.," *Journal of Big Data,* 2018.

[72] S. D. Ženko, "Is Combining Classifiers with Stacking Better than Selecting the Best One?," *Machine Learning,* vol. 54, no. 3, p. 255–273, 2004.

[73] N. Chand, P. Mishra, C. R. Krishna, E. S. Pilli and M. C. Govil, "A comparative analysis of SVM and its

stacking with other classification algorithm for intrusion detection," in *International Conference on Advances in Computing, Communication, & Automation (ICACCA)*, 2016.

[74] F. Markatopoulou, V. Mezaris, N. Pittaras and I. Patras, "Local Features and a Two-Layer Stacking Architecture for Semantic Concept Detection in Video," *IEEE Transactions on Emerging Topics in Computing,* vol. 3, no. 2, pp. 193 - 204, 2017.

[75] M. E. MichalWozniaka, "A survey of multiple classifier systems as hybrid systems," *Information Fusion,* vol. 16, pp. 3-17, 2014.

[76] J. C. a. L. Xiong, "Protein Sequence Classification with Improved Extreme Learning Machine Algorithms," *BioMed Research International,* 2014.

[77] M. M. a. Kurzynski, "On a New Method for Improving Weak Classifiers Using Bayes Metaclassifier," in *International Conference on Computer Recognition Systems*, 2017.

[78] A. S. SasanBaraka, "Fusion of multiple diverse predictors in stock market," *Information Fusion,* vol. 36, pp. 90-102, 2017.

[79] D. Zhu, "A hybrid approach for efficient ensembles," *Decision Support Systems,* vol. 48, 2010.

[80] "A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods," in *IEEE International Conference on Data Mining*, 2001.

[81] I. Czarnowski and P. Jedrzejowicz, "An approach to machine classification based on stacked generalization and instance selection," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016.

[82] G.-B. S. GaoHuanga, "Trends in extreme learning machines: A review," *Neural Networks,* vol. 61, pp. 32-48, 2015.

[83] Y. Z. A. B. K. B. P. S. B. P. Hang Chang, "Stacked Predictive Sparse Decomposition for Classification of Histology Sections," *International Journal of Computer Vision,* vol. 113, no. 1, p. 3–18, 2015.

[84] J. X. H. Z. Z. Y. J. W. Chunyong Yin, "Short text classification algorithm based on semi-supervised learning and SVM," *International Journal of Multimedia and Ubiquitous Engineering,* vol. 10, no. 12, pp. 195-206, 2015.

[85] H. P. P. P. Rekh, "Improving Accuracy of Text Classification for SMS Data," *International Journal for Scientific Research & Development,* vol. 1, no. 10, pp. 2128-2130, 2013.

[86] G. e. a. Sigletos, "Combining information extraction systems using voting and stacked generalization," *Journal of Machine Learning Research,* pp. 1751-1782, 2005.

[87] F. N. H. S. a. M. K. Koutanaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," *Journal of Retailing and Consumer Services,* vol. 27, pp. 11-23, 2015.

[88] D. E. a. J. H. H. Goldberg, "Genetic algorithms and machine learning," *Machine learning,* vol. 3, no. 2, pp. 95-99, 1988.

[89] R. a. S. P. Sikora, "Efficient genetic algorithm based data mining using feature selection with Hausdorff distance," *Information Technology and Management,* vol. 6, no. 4, pp. 315-331, 2005.

[90] D. F. R. C. T. &. M. R. L. Cook, "Combining a neural network with a genetic algorithm for process parameter optimization," *Engineering applications of artificial intelligence,* vol. 13, no. 4, pp. 391-396, 2000.

[91] S. B. E. &. D. R. Geman, "Neural networks and the bias/variance dilemma.," *Neural computation,* vol. 4, no. 1, pp. 1-58, 1992.

[92] C.-F. Tsaia, "Improving text summarization of online hotel reviews with review helpfulness and sentiment," *Tourism Management,* 2020.

[93] R. Niar, V. Jain and P. S. Nair, "Enhancing the performance measure of sentiment analysis through deep learning approach," *International Journal of Computing and Digital Systems,* 2020.

[94] Y. W. Lynne Webb, Techniques for sampling online text-based data sets, Big data management, technologies, and applications, 2013.

[95] M. R. G. T. M. M. N. C. O. a. G.-C. R. Rashid, "Completeness and consistency analysis for evolving knowledge bases," *Journal of Web Semantics,* vol. 54, pp. 48-71, 2019.

[96] G. K. J. L. S. P. D. H. S. Y. S. J. O. H. L. Jeonghun Baek, "What Is Wrong with Scene Text Recognition Model Comparisons? Dataset and Model Analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019*.

[97] J. M. B. Arauzo-Azofra, "Consistency measures for feature selection," *Article in Journal of Intelligent Information Systems,* 2008.