SENTIMENT ANALYSIS OF MOVIES USING CLASSIFICATION TECHNIQUE TO PREDICT THEIR GENRE CLASS

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

PANKAJ KUMAR

11501419

Supervisor

MR. KEWAL KRISHAN

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

Month April, Year 2017

ABSTRACT

In our busy paced life, Movies are the great source of entertainment which directly or indirectly affects our daily life. Movies have become a large business market where crores of money is invested and to recover the money ticket rates are kept high. With high ticket rates, people cannot afford to watch so many movies per month. Hence, they plan to go for certain genre of movies of their choice based on the released trailer on social networking websites as YouTube.

In our research work, we have used the Support Vector Machine (SVM) classifier using Weka 3.9 for the sentiment analysis of the movies using the subtitles of their trailer released on YouTube. We have used document level sentiment classification for the sentiment analysis of movie trailer dataset. Classification technique is used to classify the input dataset into action, drama and romance classes. We have applied Percentage Split as well as Cross Validation testing option for the comparison of classification accuracy. Percentage Split applied with SVM classifier has given the best accuracy of 83.33% compare to Cross Validation with best accuracy of 60%.We have used the Random Forest classifier to do the comparative analysis with SVM classifier. Confusion Matrix and Classification Accuracy of both the Classifier have concluded SVM as best classifier compared to Random Forest, for the given movie dataset.

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation entitled "SENTIMENT ANALYSIS OF MOVIES USING CLASSIFICATION TECHNIQUE TO PREDICT THEIR GENRE CLASS" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr. Kewal Krishan. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Pankaj Kumar

R.No. 11501419

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation entitled "**SENTIMENT ANALYSIS OF MOVIES USING CLASSIFICATION TECHNIQUE TO PREDICT THEIR GENRE CLASS"**, submitted by **Pankaj Kumar** at **Lovely Professional University, Phagwara, India** is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

Mr. Kewal Krishan

Date:

Counter Signed by:

1) **Concerned HOD:**
   HoD's Signature: _____

   HoD Name: _____

   Date: _____
2) **Neutral Examiners:**

   External Examiner

   Signature: _____

   Name: _____

   Affiliation: _____

   Date: _____

   Internal Examiner

   Signature: _____

   Name: _____

   Date: _____

ACKNOWLEDGEMENT

Working on the research work was like a journey where I have gone through various phases of research work. I have worked hard to achieve the desired result for my research work. However, it would not have been possible without the kind support and help of my mentor Mr. Kewal Krishan. I would like to extend my sincere thanks towards my mentor for his constant support and the invaluable supervision.

I would like to express my gratitude towards Mr. Dalwinder Singh, HOD for their constant support and insightful suggestion during the journey of my research work.

I would also like to express my gratitude towards Dr. Rajeev Sobti, HOS for their kind co-operation and encouragement.

I have my family, friends, colleagues and well-wishers on-board for this journey to make it successful. I would like to express my heartfelt gratitude to each one of them for their invaluable suggestions and guidance.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES

LIST OF TABLES

Checklist for Dissertation-II Supervisor

Name: _____ UID: _____ Domain: _____
Registration No: _____Name of student:_____

Title of Dissertation:

_____

- ☐ Front pages are as per the format.
- ☐ Topic on the PAC form and title page are same.
- ☐ Front page numbers are in roman and for report, it is like 1, 2, 3…….
- ☐ TOC, List of Figures, etc. are matching with the actual page numbers in the report.
- ☐ Font, Font Size, Margins, line Spacing, Alignment, etc. are as per the guidelines.
- ☐ Color prints are used for images and implementation snapshots.
- ☐ Captions and citations are provided for all the figures, tables etc. and are numbered and center aligned.
- ☐ All the equations used in the report are numbered.
- ☐ Citations are provided for all the references.
- ☐ **Objectives are clearly defined.**
- ☐ Minimum total number of pages of report is 50.
- ☐ Minimum references in report are 30.

Here by, I declare that I had verified the above mentioned points in the final dissertation report.

Signature of Supervisor with UID

CHAPTER 1
INTRODUCTION

Sentiment Analysis is the most happening and trending research area which are applicable to various fields as stock marketing, weather forecasting, social networking and e-commerce business world. Growing importance of sentiment analysis is due to the popularity of various means of communication on public platforms as social media, print media and mass communication media.

Sentiment does not consist of facts however it provides subjective impressions. Sentiment means:

- Feeling
- Opinion
- Emotion

Sentiment Analysis is the analysis of the person's behavior, opinion, their attitude which is going to predict the polarity of that sentiment as well as their future moves. Sentiment Analysis and Opinion mining are synonyms in nature as they express the same meaning. However, some peoples have different view point about sentiment analysis and Opinion mining.

Opinion mining analyzes the person's opinion about any entity whereas Sentiment Analysis classifies the sentiment expressed in a text and predicts the polarity of sentiment. So, Sentiment analysis detects the sentiment expressed in a text by different text mining techniques and predicts the polarity of sentiment.

1.1 SENTIMENT ANALYSIS TECHNIQUES

Sentiment Analysis is the approach for analyzing the behavior of a person's sentiment based on dataset which are collected from various means of social media. Social media is a platform which provides the opportunity for peoples on global platform, to interact with each other, shared their views about different products, movies, software etc. With the wide reach of internet, social media becomes the prominent and easy to communicate with different types of people and can provide their opinion on various issues.

Various prominent companies are using the review given by their customer regarding their various products in the form of comments to improve the quality as well as variety of their products for better customer satisfaction.

Companies such as Amazon, Flipkart, and Snapdeal are analyzing the shopping behavior to provide them the recommended product which they are looking for. Twitter has provided the twitter API in their development section to extract the twitter data of various users using your own account for the sentiment analysis.

Sentiment Analysis approach mainly consist of two types as shown in the figure 1.1

- Machine Learning based approach
- Lexicon based approach

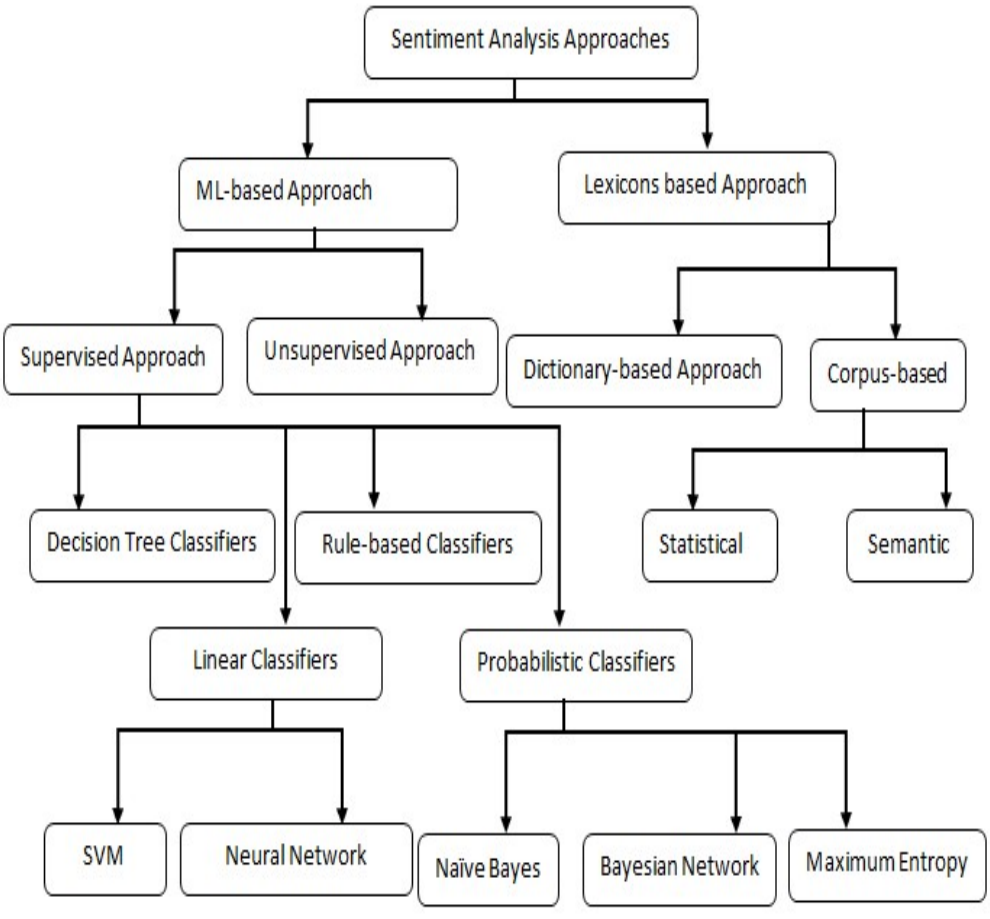

Figure 1.1: Sentiment Classification Techniques

Machine Learning based approach is a type of artificial intelligence based approach where it provides the ability to learn without being overtly programmed. It provides the features for development of computer program that can modify when open to fresh data. It is of two types:

- Supervised based approach
- Unsupervised based approach

In case of *Supervised based learning approach,* inputs are given with their labels defined in the form of training dataset and their desired output are also defined. The fresh data or the testing dataset are evaluated based on the learning of classifiers through training dataset and classify them into various labels. Classification technique is the based example of Supervised based learning approach.

Supervised based learning approach consist of following classifiers:

1. Decision Tree Classifier
2. Rule Based Classifier
3. Line Classifier
    - SVM (Support Vector Machine)
    - Neural Network
4. Probabilistic Classifier
    - Naïve Bayes
    - Bayesian Network
    - Maximum Entropy

In case of *Unsupervised based learning approach,* there are no labels defined for the learning algorithm and it finds its structure by itself. It is a self learning approach where a structure is formulated based on some hidden pattern or may be grouping of data and hence data are classified into various labels based on the learning pattern.

Clustering is the best example of *Unsupervised based learning approach.* Clustering is the Unsupervised based learning approach which finds the cluster of data objects which are similar in their pattern. Members of same clusters are more like each other than the members of other clusters. In case of Clustering, data are segmented into various clusters based on similarity of their pattern.

In case of Classification which is the example of Supervised based learning approach, data are segmented into various defined lacels whereas in case of Clustering, data are segmented into various labels or clusters which are not previously defined. An example of Clustering is shown in the figure 1.2.

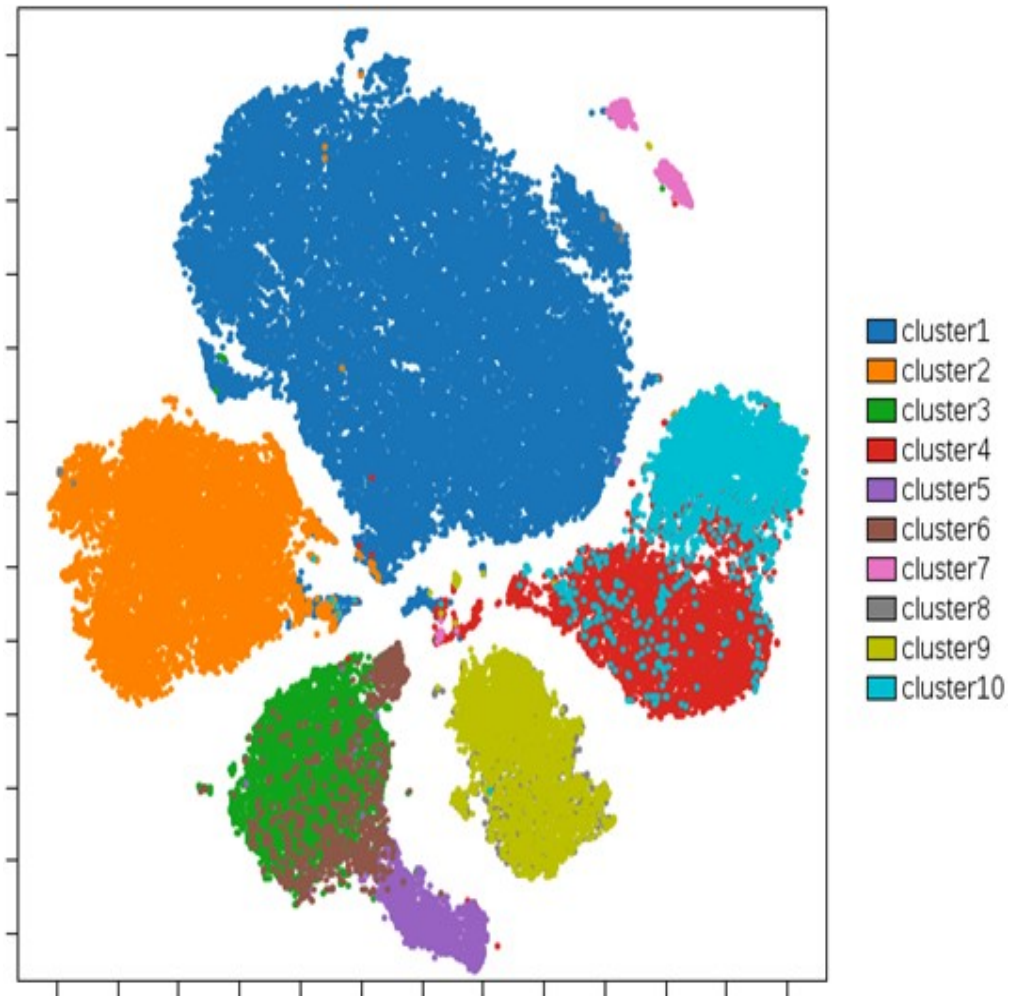

Figure 1.2: An example of Clustering

Lexicon based approach is used to extract sentiment from the text data. This approach uses semantic orientation calculator which is going to provide a degree of opinion and subjectivity in text. Semantic orientation uses dictionary of words which has well defined semantic polarity for each word to calculate the semantic orientation measure or score. Dictionary can be created manually or automatically using seed words to expand a list of words. Most of the lexicon based approach research has engrossed towards using the adjectives as pointers of semantic orientation of text.

For creating a dictionary, first a collection of adjectives and their corresponding semantic orientation values are compiled into a dictionary. Then for any given text dataset, all adjective are extracted and marked with their semantic orientation value using the predefined dictionary score. Semantic orientation score is accumulated to provide the polarity score for the text for each orientation label. Lexicon based approach is best explained using the figure 1.3.

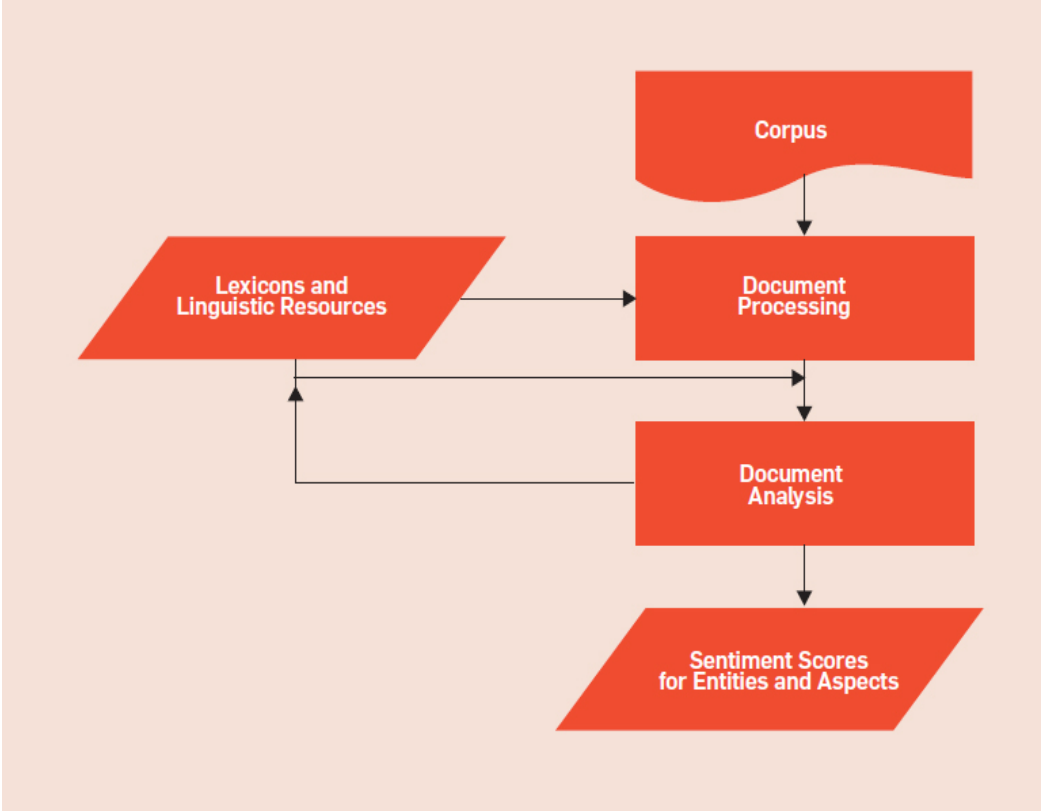

Figure 1.3: An example of Lexicon based approach

1.2 SENTIMENT CLASSIFICATION

Sentiment Analysis is studied and implemented at different levels such as:

1. Document level classification
2. Sentence level classification
3. Feature level classification

Sentiment Analysis for a given dataset is done by one of the above sentiment classification level.

Each of the sentiment classification level has their properties to classify the dataset into various labels. It all depends on the type of input dataset to determine which level of sentiment classification will provide better accuracy and performance.

1. *Document level classification*: In this level of sentiment classification, each document is treated as a single information unit. That means each document is treated as a single instance and a label is a collection of instances.

   For example, when we are doing the sentiment analysis of movie using the transcript of the trailer, data is collected on document level providing a single document for the single movie transcript. And labels such as *action*, *romance* and *drama* are given to folders which has the collection of all action, romance and drama genre movies in each folder respectively.

   On the given dataset the classification techniques are applied using various classifiers to classify the various movies, which are the instances in document form into various labels.

2. *Sentence level classification:* In this level of sentiment classification, each sentence is treated as a single information unit. Analysis is done to predict the polarity of sentiment or emotion expressed in a sentence. Sentences are fall into subjective and objective type classes.

   Subjective type class is a collection of different types of sentiments of input sentences which are going to be extracted in the sentiment analysis phase. *Objective type class* is a collection of neutral sentiment sets.

   To improve the performance and accuracy of the classification, it is suggested to remove the objective sentences from dataset before classify the polarity of dataset.

3. *Feature level classification:* In this level of sentiment classification, various aspects of text dataset are used to classify the dataset into various labels. Different feature [12] can derive different sentiment for the same entity. For example, a cell phone can have better display feature and a low battery life. This task involves various phases such as identification of object features, determining opinion orientations, grouping synonyms such that they are able to extract different aspect or features for a particular entity. And as a result different sentiments for different aspect of single entity can classified into various labels.

1.3 SENTIMENT ANALYSIS PHASES

Sentiment Analysis for a given input dataset consist of various phases, where data is processed at each phases to classify the data into various label as shown in the figure 1.4.

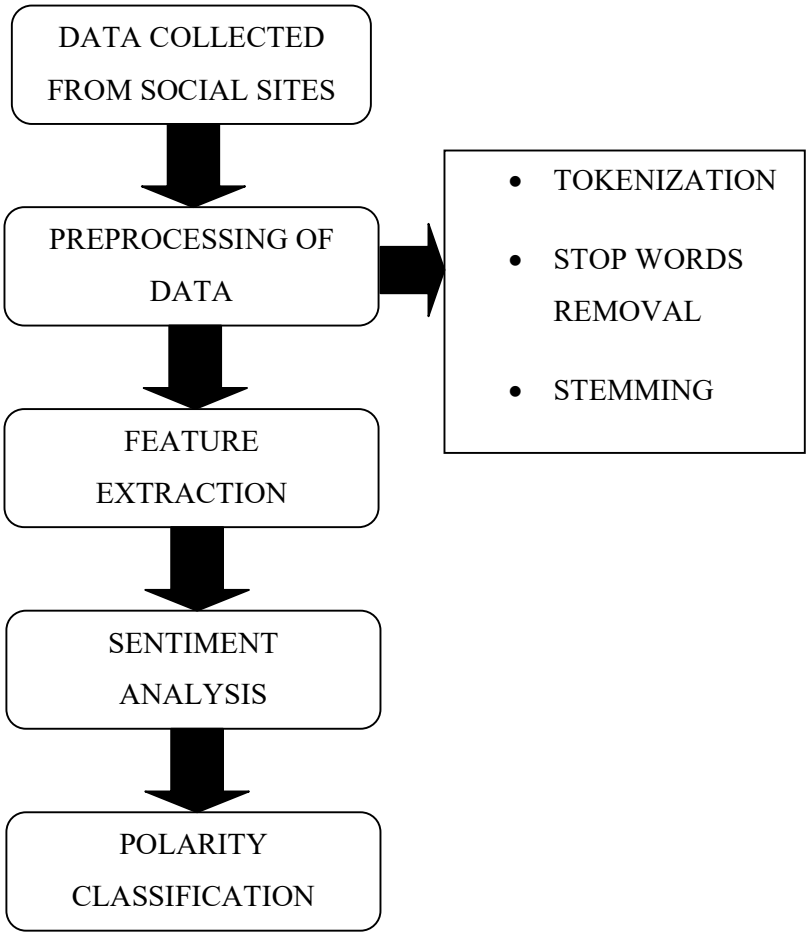

Figure 1.4: Various phases of sentiment analysis

First Phase consist of *Data Collection,* where data is collected from various social media sites as per the requirement by the researcher for sentiment analysis. Data is collected mainly from social sites because social sites provides a platform on global level, so that users can interact with each other, provide their opinion about various products such as movies, software, products used in daily life etc.

To generate such huge amount of data from billions of users can not be possible without use of social media. A pictorial representation of social media generally used for sentiment analysis is shown in the figure 1.5.

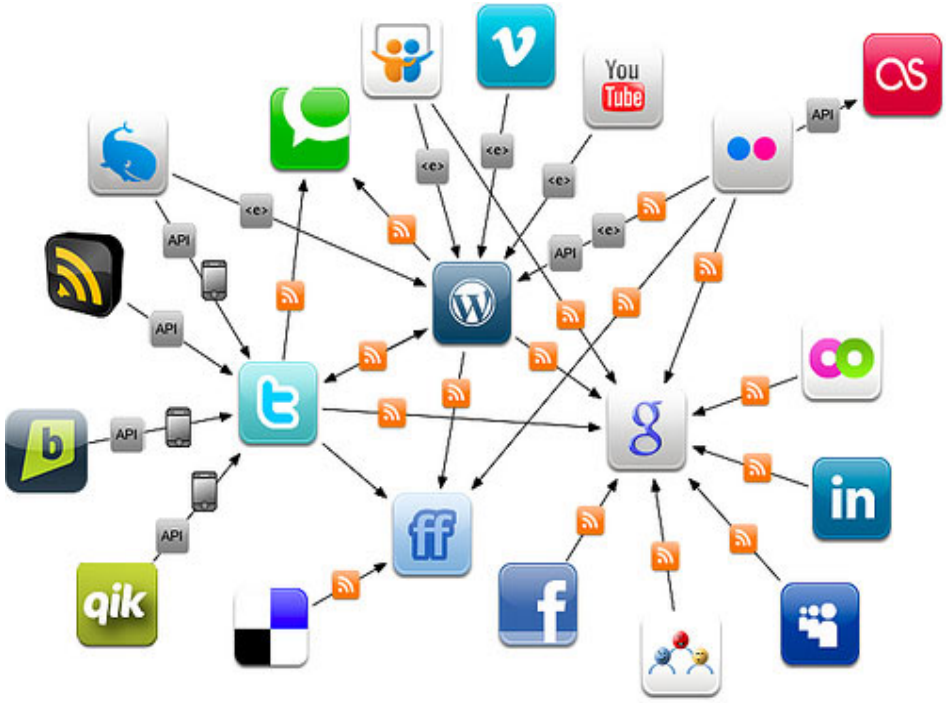

Figure 1.4: Various sources of social media data

Second Phase consist of *Preprocessing of data,* where data is processed using various tools before using it for analysis purpose. Since data are collected in raw form which are not suitable for the analysis purpose. Following are the mainly used preprocessing tools as:

- Tokenizer
- StopWordsHandler
- Stemmer

Tokenizer is mainly used to split the text strings into word tokens, which is of various types having different properties and features.

StopWordsHandler is used to identify and remove the stop words from the input dataset.

Stemmer is used to find out the root word and stem out the other versions of same word. For example, fly, flying, flies having same root word i.e. fly.

Third phase consist of *Feature Extraction,* where various features or aspect of same entity are extracted to classify them into various labels based on different emotional aspect for same entity. This phase is mainly used in the case of Lexicon based approach for sentiment analysis.

Fourth phase consist of *Sentiment Analysis,* where different classifiers or clusters are applied on dataset based on which type of machine learning the researcher is using to classify the data into various labels.

Fifth and last phase consist of *Polarity classification and Visualization,* where data are classified into different labels or polarity and result are visualized through graph, trees and confusion matrix.

1.4 SENTIMENT ANALYSIS APPLICATIONS

- To predict market result based on sentiment analysis of news, blogs and social media data.
- To identify and analyze Fake news.
- To compute customers satisfaction level by analyzing their reviews.
- To provide items recommendations to customers on E-commerce website.
- Voting advice application using the sentiments of the voters [4].

CHAPTER 2
REVIEW OF LITERATURE

"Benchmarking Twitter Sentiment Analysis Tools" [1] presents findings of a detailed analysis of twitter sentiment analysis tools. Here, 20 tools have been examined for sentiment polarity classification performance which include commercial as well as freely available tools. 20 tools (15 stand-alone commercial tools and 5 workbench tools) were applied across 5 test beds i.e. Pharma, Retail, Security, Tech, Telco and their polarity efficiency were compared.

"New Words Enlightened Sentiment Analysis in Social Media" [2] proposed new words based sentiment analysis methods. NWLB was a improved lexicon based method. NWLB worked fine for the contents having new words which are not effective for all the contents. So, NWSA was proposed which was the combination of NWLB and MLBM through an ensemble learning way. NWSA was effective for normal content as well as for content containing new words.

"Combining classification and clustering for tweet sentiment analysis" [3] proposed an algorithm which will combine cluster and classifiers for better accuracy. Algorithm used the result provided by clusters to refine the tweet classification based on assumption that similar instances from same cluster had same class label. Algorithm which was capable to use classifier as well as cluster collective to improve the classification accuracy.

"Approaches, tools and applications for sentiment analysis implementation" [4] discussed different sentiment classification approaches, various sentiment analysis tools and their advantages amd limitations. This paper discussed various sentiment analysis tools which are used with different techniques used for sentiment analysis. Various fields of sentiment analysis application had been discussed as Marketing, finance, politics etc.

"Sentiment Analysis Using Harn Algorithm" [5] proposed an algorithm known as HARN algorithm which could correctly classify the input satatement based on domain

in which the word was being used.

It was an supervised learning method which was introduced to solve one of the challenges of sentiment analysis i.e. domain specific adaptation. This method used domain dictionary, basic structure of sentences and pre-defined polarities to classify the given input sentence.

“Opinion Mining and Sentiment Analysis” [6] proposed a new algorithm to determine the effects of an average person’s tweet over fluctuation of stock prices of a particular company. Algorithm was used for analysis of twitter data thus providing different polarity of sentimental analysis. Efficiency of the result was measured in terms of accuracy rate and time complexity.

“AffinityFinder : A System for Deriving Hidden Affinity Relationships on Twitter Utilizing Sentiment Analysis” [7] proposed a system known as “AffinityFinder” which derived a potential friendship relationship among twitter users. Every tweet ran through the sentiment analyzer word by word. Affinity graph was build using the affinity score calculated during analysis process as shown in the figure 2.1

Figure 2.1: Affinity Graph

AffinityFinder was the system built for deriving hidden affinity relationship amongst twitter users. The proposed system collected and analysed the tweets of various users to derive the relationship scores that reflected the affinity degrees among twitter users.

Twitter data was collected and stored using MongoDB database.Analysis module used distributed data which were processed among various nodes providing the result in the form of affinity score. Affinity score was used to draw the affinity graph. A detail architecture of AffinityFinder is shown in the figure 2.2.

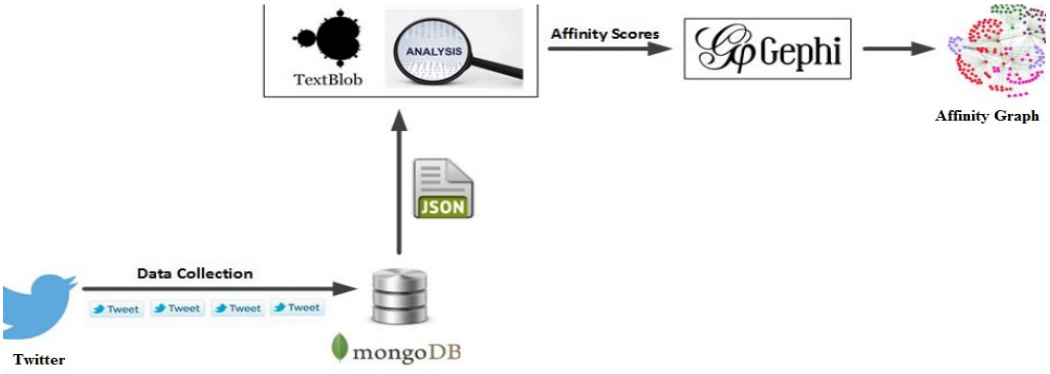

Figure 2.2: Architecture of AffinityFinder

"TSentiment: On gamifying Twitter sentiment analysis" [8] proposed a game with a purpose known as "TSentiment" that was used the compute feature of human being which were tough and difficult task for the computer system. A internet based game was created using more than 75000 tweets collected from various sources and students needed to play that game. Result provided by this analysis that it was useful for the scenario where decision were made based on person's feeling.

"Comparative Study of Classification Algorithms used in Sentiment Analysis" [9] described a comparative study of various classification algorithm such as Naïve Bayes, Max Entropy, Boosted Trees, Random Forest used in sentiment analysis. Use of a particular algorithm should be dependent on kind of particular input provided by the user. Study provided that every kind of classification model had its own benefits and drawbacks.

"Affective-feature-based sentiment analysis using SVM classifier" [10] presents a model which was used for representing the input text to solve the data dispersion problem.

SVM classifier was used which worked with the result of training set to classify the testing input data.Various critical issues of sentiment classification were verified by experiments such as determining the method for selecting the features and selecting the feature dimension.

“Sentiment analysis: A multifaceted problem” [11] discussed Sentiment analysis as a multifaceted problems which contain many sub-problems. This article discussed about multiple challenges of sentiment analysis such as identification of various objects, extracting the features, classification of opinion polarity etc.

“Sentiment analysis algorithms and applications: A survey” [12] discussed a survey details of sentiment based analysis which were recently updated. This survey paper tried to provide nearly full image of Sentiment Analysis techniques and the related field with brief details. Many recently proposed algorithms and various Sentiment Analysis applications were investigated and presented briefly in this survey.

“Multimodal Sentiment analysis of spanish online videos” [13] addressed the job of presenting a method that integrate different mode of communication such as audio, video and textual properties for identifying the sentiment from online videos. Dataset was created by collecting various spanish videos from social networking sites as youtube and analysis of the newly created dataset was done to determine the sentiment polarity. A comparative study was done between simple text file sentiment analysis and analysis of dataset consisting multi mode data.

“Online Analysis of Sentiment on Twitter” [14] examined the previous works on sentiment analysis on twitter and proposed a new model to increase the efficiency. In the proposed model, high speed dataset were used and sentiment analysis algorithm were used to predict the sentiment of the person in a timely manner without any delay.

“An improved sentiment analysis of online movie reviews based on clustering for box-office prediction” [15] discussed the various approaches and methods of sentiment analysis to analyse the relationship between review of movies and their box office performance. Clustering was used aling with classification model to increase the accuracy level of classification. This work showed a regressive model which were automate in nature having very good accuracy was able to predict the box-office sale.

"A Sentimental Education: Sentiment Analysis using Subjectivity Summation based on Minimum Cuts" [16] proposed a sentiment analysis method which was applied on the subjective class of the input text document. This work was able to identify the relation between the sentiment polarity classification and the identification of subjective nature of the text document. Identification of subjective nature of document was able to compress the input text reviews in shorter form that still contain the polarity information.

"Sampling techniques for streaming dataset using sentiment analysis" [17] proposed the model which used the sentiment analysis techniques on buffering dataset sample to analyse the polarity of input dataset. Sampling techniques were applied to extract out a sample of streaming dataset. Sample streaming dataset of twitter were analysed to find out the polarity of the sample dataset.

"Multiclass Classification and class based Sentiment Analysis For Hindi Language" [18] proposed the model which classified the sentiments of the document consisting of hindi language text. Sentiment analysis was done using HSWN and LMC classifier to predict the polarity of each class. A comparative study for better accuracy was done between sentiment analysis with only HSWN and sentiment analysis with combination of HSWN and LMC classifier.

"Sentiment analysis: A combined approach" [19] proposed a new combined method which was a combination of different approaches of sentiment analysis classification. Proposed approach was applied on the dataset from various sources consisting of different types of reviews. Result showed that combined approach was providing more accuracy than a single approach of sentiment analysis.

"Sentiment analysis of English Tweets using RapidMiner" [20] used sentiment analysis techniques to classify the twitter dataset which was collected from various twitter users into various polarity using RapidMiner tool. Collected dataset were pre-processed before uploading the dataset in RapidMiner. Classifier were trained using training dataset and applied on testing dataset to predict the polarity of tweets.

"Sentiment trend analysis in social web environments" [21] proposed a method for sentiment analysis of trends on social web media using ACO (Ant Colony Optimization) algorithm and SentiWordNet.

Data was collected from *feeling* tag of each user's Facebook wall in the form of Resource Description Framework (RDF) triples. Algorithm was used to extract the trend list for each user's sentiment from social media platform such as Facebook. Sentiment score (positive and negative) were applied using SentiWordNet for each user and a total sentiment score were computed. Based on the sentiment score online trend were analyzed and validated with trends of real daily life on social media.

"Sentiment analysis of Twitter data: Case study on digital India" [22] provided a case study on digital India campaign with sentiment analysis on social media data such as twitter data. Data was collected from twitter using twitter API with Python code. Dictionary based approach was applied for the sentiment analysis of twitter data using NLTK modules. Data collected from twitter were matched with positive and negative words of Dictionary and classified accordingly. Remaining words were classified as neutral.

"Sentiment analysis of movie reviews: A study on feature selection & classification algorithms" [23] provided sentiment analysis of movie review using the data provided by the various users on IMDB web site through comment sections for various movies. Movie review comment expressions were examined to classify the polarity of movie review on a scale of 0 to 4. Feature extraction was applied and those features were used to train the classifier to classify the movie review comments into correct label. A comparative study for various different classifiers providing different accuracy was performed.

"Feature level Sentiment Analysis on Movie Reviews" [24] proposed a model which classified the movie review dataset into positive and negative classes based on the features extracted from the dataset by handling synonyms, negation and conjunction with appropriate pre-processing steps. Scores of the movie review were calculated with the help of SentiWordNet tool. Sentiment score obtained were used to classify the movie review dataset into positive and negative classes according to the score.

"Sentiment Analysis for Movie Reviews" [25] proposed a model which was used for sentiment analysis of movie review to classify the movie review into positive and negative classes. Here SKLearn module has been used as a tool for the sentiment analysis of the movie review dataset into positive and negative classes.

Three feature extraction methods were used to extract the features from the dataset and used that features to train the classifiers and as a result trained classifiers classify the data into positive and negative classes. Feature extraction methods were *Bag of words*, *N-Gram Modeling, TF-IDF Modeling.* Various classifiers such as Naïve Bayes, Random Forest, KNN, and SGD were used providing different accuracy. Random Forest provided the best accuracy for the given movie review dataset.

"Sentiment Analysis of Call Centre Audio Conversations using Text Classification," [26] proposed a model which worked in two phases. In first phase, we auto generate text copy from the audios using speech recognition technology and in next phase we were analyzing the text transcript copy using sentiment analysis technologies as shown in the figure below.

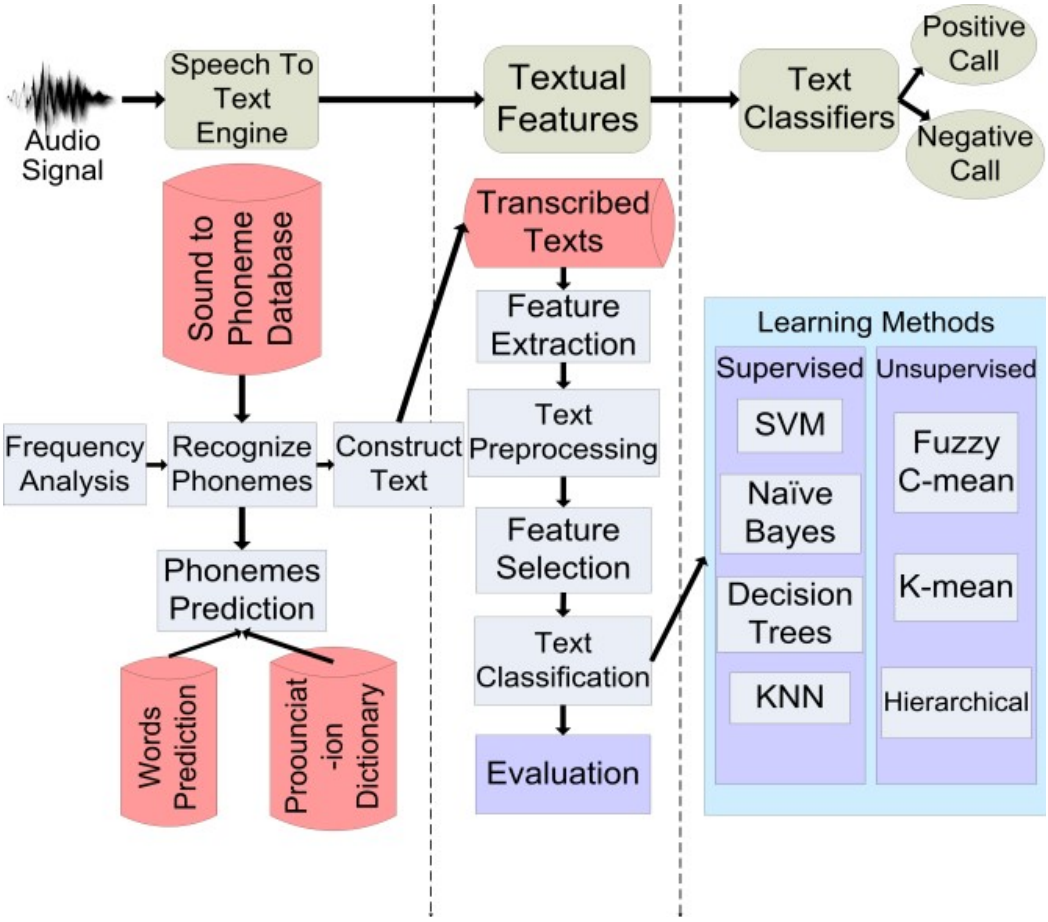

Figure 2.3: Text based classification for Call Centre Conversation

Various classifiers were used for sentiment analysis of the text data which classified them into positive and negative labels with different accuracy. Experiment provided

that speech engine is gender dependent having better accuracy for female conversation as compare to male conversation.

"Sentiment analysis using product review data" [27] proposed a common process of sentiment analysis to classify the online product review dataset into positive and negative classes with detail process description. Data collected for the analysis were from the *amazon.com*. A rating system was applied for the online product review which was provided by various customers as shown in the figure.

Star Level	General Meaning
★	I hate it.
★★	I don't like it.
★★★	It's okay.
★★★★	I like it.
★★★★★	I love it.

Figure 2.4: Rating System for Products

Here sentiment analysis was done on sentence level. Sentiment Analysis was done in three phases as shown in the figure.

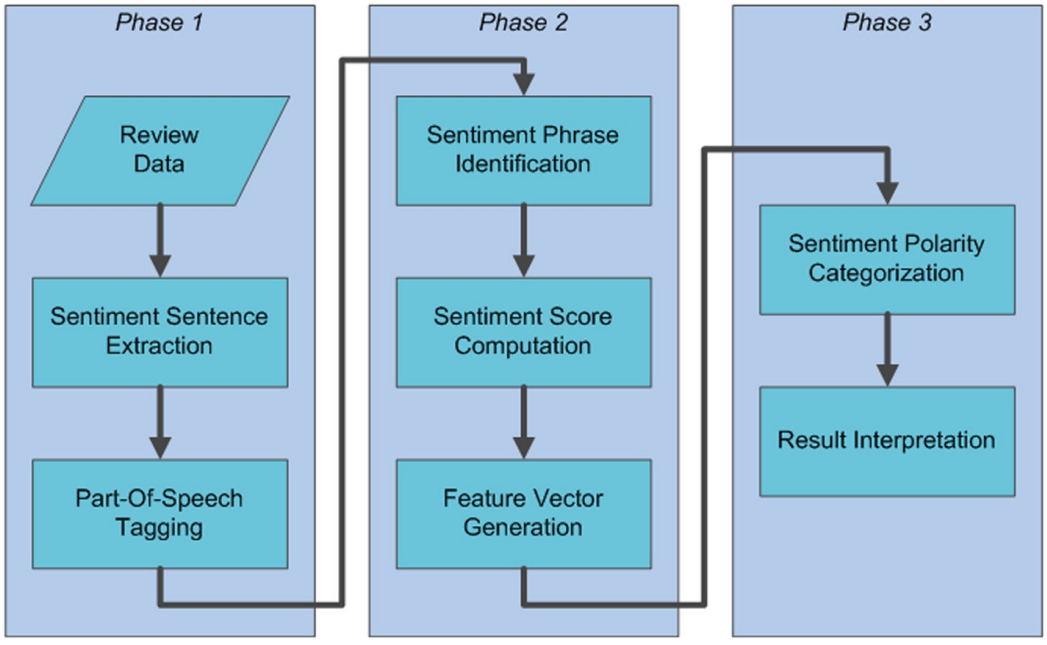

Figure 2.5: Sentiment polarity Classification processes

In first phase, Sentiment sentence were extracted from the product review dataset. In second phase, for negation phrase identification, an algorithm was proposed and implemented. Sentiment score was computed using SentiWordNet tool. In third and last phase, experiment was conducted to classify the sentiment expression into positive and negative classes. Performance of these classification models were compared and evaluated.

"A Literature Review on Opinion Mining and Sentiment Analysis" [28] presented a synopsis on Sentiment Analysis or Opinion Mining. Various methods, tools and dataset used by the researcher with their accuracy were discussed in this review paper. Sentiment Analysis are done on three levels i.e. document level, sentence level and feature level for the classification of dataset into different classes. There are mainly two approaches of sentiment analysis i.e. Machine Learning based approach and Lexicon based approach. Supervised and unsupervised are the two types of Machine learning approach. Dictionary and Corpus based approach are two types of Lexicon based approach. There are various classifiers which were used to classify the dataset into various classes depending on the type of data.

"Real time sentiment analysis of tweets using Naive Bayes" [29] contained the implementation of Naïve Bayes classifier to train data using twitter database. A method was proposed to improve classification. SentiWordNet tool was used along with Naïve Bayes classifier to improve the accuracy for the classification of tweets. SentiWordNet was used to calculate the sentiment score which was used to classify the dataset into various classes. Python was used along with NLTK module for the implementation of proposed model. Python twitter API was used to extract data from the twitter media.

"Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis" [30] discussed the effect of sentiment analysis pre-processing step on performance of sentiment classification for given dataset. Research summed up the performance of sentiment classification with 6 pre-processing methods which were using 2 feature models and 4 classifiers on 5 twitter datasets. Experiment showed that accuracy were improved when using pre-processing methods when replacing negation, expanding acronyms as compared to removing numbers, URLs and stop words.

CHAPTER 3
PRESENT WORK

This chapter is divided into three sub sections which consists of Problem Formulation, Objectives of the Study and Research Methodology.

3.1 PROBLEM FORMULATION

In our research work, we have focused on the Sentiment Analysis of Movie dataset using the transcript or subtitle of the movie trailer to predict the genre of the movie. We have used the supervised learning approach using classifiers to predict the polarity towards different movie genre as Action, Romance and Drama.

Before the release of any movie, their trailer is released on the YouTube social networking website. So for our research work where we have worked on the movie dataset have collected from the YouTube website.

IMDB (Internet Movie Database) website www.imdb.com has recognized a total of 20 movie genre. However, we have observed that some of the movies genre shows a high correlation. So we have made a total of 3 labels for 20 movie genre as shown in the table 3.1.

Table 3.1: Movie Genre distribution among labels

LABEL	GENRE INCLUDED
ACTION	THRILLER, HORROR, ACTION, CRIME, WAR, ADVENTURE, SCI-FI, MYSTRY, FANTASY
DRAMA	FAMILY, COMEDY , HISTORY, SPORTS, SHORT MOVIE, DOCUMENTRY, BIOGRAPHY, DRAMA
ROMANCE	ROMANTIC, MUSICAL

We have applied the machine learning approach in the form of supervised based learning approach, where classifiers is used to classify the movie trailer dataset into three movie genre labels as *action, drama, romance*.

We have worked on *document level based sentiment classification*, where we have collected data in the form of document and all the documents are stored in related

movie genre label as *action, drama, romance*. We have used the movie Wikipedia as a reference to store the movies in respective movie genre label for released movies. And for unreleased movie we have analyzed the movie trailer subtitle in text document to store in the respective movie genre label based on the analysis of released movie trailer.

We have used the two classifiers, SVM (Support Vector Machine) and Random Forest to classify the movie trailer dataset into three labels such as *action, drama and romance.* Result obtained in the both Classifiers have analyzed for better accuracy and performance. Hence, providing the result of best Classifier for the given movie trailer dataset.

3.2 OBJECTIVES OF THE STUDY

In our research work we have worked to achieve the following objectives using the sentiment analysis techniques:

- We have classified the movie trailer dataset into three movie genre labels such as action, drama and romance. Movie trailer dataset has been collected in the form of document to fulfil the need of document based sentiment classification of movie trailer.
- A comparison based experiment has been done for the two classifiers i.e. SVM and Random Forest to find out the best classifier providing best accuracy and performance for our given dataset.
- A comparison based experiment has been done for the testing options i.e. Cross Validation and Percentage Split to find out the better way of testing the data to provide higher accuracy and performance for our given dataset.
- We have conducted the experiment to find out the best *confusion matrix* (which describes the performance of a classifier) on what range of percentage split or cross validation. Confusion Matrix has provided the label classification result consisting rows for actual values and columns consisting predicted values.
- We have come to know the variation between predicted and actual values using confusion matrix, when there has been variation in the testing options such as Percentage Split or Cross Validation.

3.3 RESEARCH METHODOLOGY

This section consists of *various phases of methodology*, *Flow diagram of experiment* and *Development/Analysis tool*.

3.3.1 VARIOUS PHASES OF METHODOLOGY

In our research work, we have the following phases:

1. Problem Formulation
2. Data Collection
3. Preprocessing of data
4. Classify the data
5. Analysis and Visualization of the result

In first phase, problem has been formulated for our research work. In our research work, we have proposed to do the sentiment analysis of a movie using the transcript or subtitle of the movie trailer to classify the movies into three movie genre label as *action, drama* and *romance*. To choose the best classifier for our dataset, we have proposed to do the comparison based analysis between two best available classifiers i.e. SVM (Support Vector Machine) and Random Forest, having best available accuracy as known from the literature survey.

In second phase, data has been collected from the social networking website YouTube as shown in the figure 3.1, having the website address www.youtube.com.

Figure 3.1: Data collected from social networking website YouTube

We have extracted the data of a particular movie in a single text document file as we were working on the document level sentiment classification. There are various ways to extract the subtitle or transcript data for a particular movie trailer. Following are some ways of extracting subtitle data from YouTube:

- Play the video. If the video have the *cc* icon, then the video consist of subtitle for the video. Go to *more* section, which is below the playing video. Select *Transcript,* the transcript for that particular video will be generated. We can copy the data and store it into text document.
- Using Developer tool
- Using online website which generate the subtitle file on submission of video file.
- Using Subtitle Extractor tools etc.

We have applied the first method from the above mentioned way to extract the data, by selecting the *transcript* from the *more* section of YouTube video. Data has been copied and stored in text document which has been kept under respective movie genre label. Three folders have been created for three movie genre label as action, drama and romance. Each text document has been the representation of each movie which has been stored into their respective label.

In third phase, Preprocessing of data has been done because the data we extracted were in raw form which was not suitable for the sentiment analysis to provide better accuracy. Since we have stored the data in text format, our analysis tool (WEKA) has failed to read in text format. However, Weka has provided the inbuilt converter such as *TextDirectoryLoader* which will load the dataset in text directory format. We have applied some filtering mechanism to preprocess the recently loaded dataset which has become suitable for applying sentiment analysis techniques. Following were the filtering tool used in our research work:

- Tokenizer
- StopWordsHandler
- Stemmer

Tokenizer is mainly used to split the strings of the text document into word tokens.

In our research work we have applied *WordTokenizer,* which is using java to split a string of the text document into words. The whole strings of the dataset have been split into words. Each word has been counted as an attribute.

StopWordsHandler is used to identify and remove the stop words from the input dataset. In our research work we have used *Null* StopWordsHandler which does not remove any stop words.

Stemmer is used to find out the root word and stem out the other versions of same word. In our research work we have used *NullStemmer* which is a dummy stemmer that performs no stemming at all.

In fourth phase, Classifiers have been applied to classify the movie into three defined labels as action, drama and romance. In our research work we have used the following two classifiers:

1. *Support Vector Machine (SVM):* It is supervised based learning algorithm which is mainly used for classification. Here each data item is plotted as a point in n-dimensional space, where n is the number of features we have. Classification is done to find the hyperplane which clearly differentiate the two classes. SVM classifier can be easily understood with the help of figure 3.2.

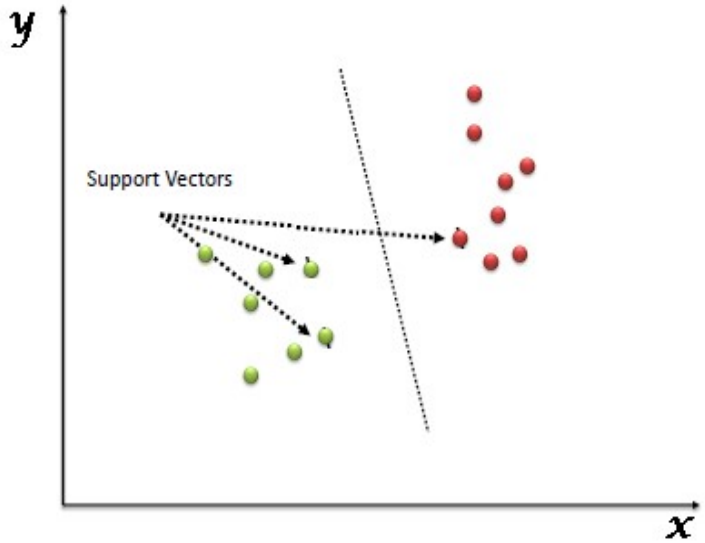

Figure 3.2: An example of Support Vector Machine

2. *Random Forest (RF)*: It is a versatile supervised based learning algorithm which has the capability to perform classification as well as regression. It is an ensemble learning model where a list of fragile models is combined to form a powerful model. Working architecture of Random Forest classifier is shown in the figure 3.3.

Figure 3.3: Working architecture of Random Forest

Working: Sample is made with N cases of training set taken at random but with replacement. This sample is used as training set for growing the trees. If there is K input variable then k variables are selected at random from the large set K. Split of node is node by best split on k variables. While growing the tree k value is kept constant. Each tree is grown to their largest extent without pruning. Predicate new data by aggregating the prediction as taking majority of votes.

In Classification phase, we have selected the classifier and tested the input dataset with two of the following testing options as:

- Use training set
- Supplied test set
- Cross Validation
- Percentage Split

We have applied *Percentage Split and Cross Validation* for comparison of better accuracy. We have also conducted a comparison based experiment to test both the classifier using theses testing option.

In fifth and last phase, we have done the analysis of result using Classification accuracy, TP rate, FP rate, Precision, Recall, F-measure, confusion matrix and various visualization measures such as Cost/Benefit analysis, threshold curve, margin curve and cost curve. Classification of movie into various label have been represented through *Confusion Matrix.*

3.3.2 FLOW DIAGRAM OF EXPERIMENT

Flow diagram for sentiment analysis of movie using the subtitle of the movie trailer consists of following steps:

Step 1: *Collection of data from the social networking website YouTube*

Data has been collected from social networking site YouTube and stored in the text document format. Three folders consisting of labels action, drama and romance have been defined which kept the respective movies documents.

Step 2: *Preprocessing of Data*

Extracted data has been preprocessed using various filtering tools such as WordTokenizer, NullStemmer and Null StopWordsHandler.

Step 3: *Classify the data*

In this step, various classifiers such as SVM and Random Forest have been selected and dataset has been tested using two testing options (Percentage Split and Cross Validation) one by one.

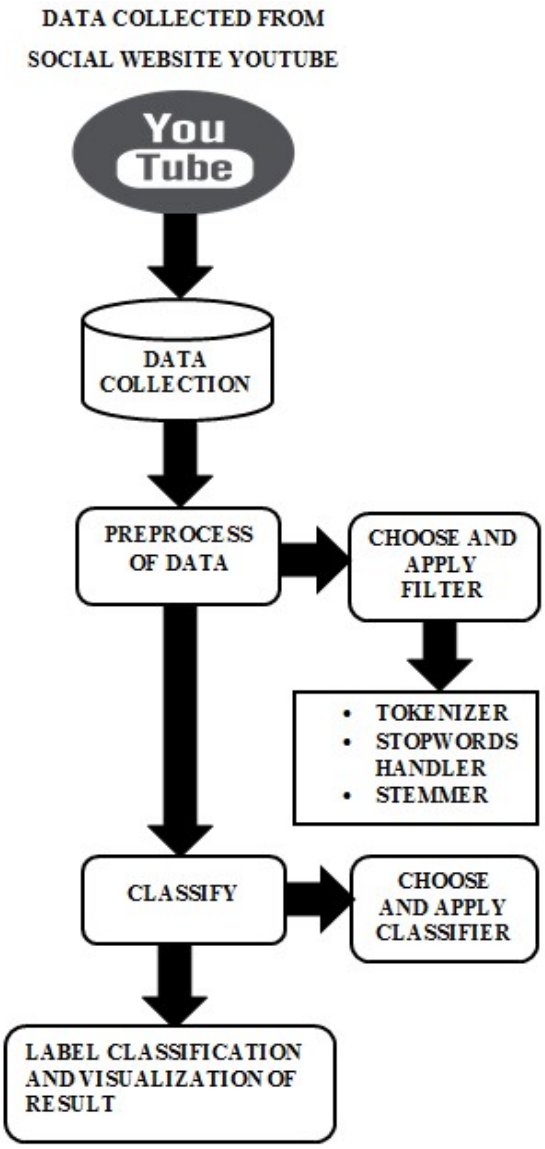

Figure 3.4: Flow diagram for sentiment analysis of movie

Step 4: *Label Classification and Visualization of Result*

In this step, dataset has been classified into three defined labels such as action, drama and romance. Correctly classified movie instances have been showed by the classification accuracy. Confusion Matrix has been shown as a result which showed the performance of the selected classifier. Confusion matrix showed the matrix of actual instances with predicted instances. Various visualization measures such as Cost/Benefit analysis, threshold curve, margin curve and cost curve.

3.3.3 DEVELOPMENT/ANALYSIS TOOL

WEKA is a data mining system tool which implements data mining algorithm, has been developed in the year 1993 by the University of Waikato, New Zealand. Weka has been the ultra-modern facility for developing the Machine Learning techniques and applications for the vast real world problems. It consists of the entire data mining algorithm used in Machine learning based approach. Data mining algorithm which has been written by the system can be applied directly to analyze the accuracy of the various data mining concepts such as Classification, Clustering etc. In our research, we have used WEKA 3.9 for sentiment analysis of our movie dataset.

Weka tools provide the facility to connect with various language based development tools such as Eclipse through JDBC/ODBC. Weka implements the algorithm for the following data mining concepts as:

- Data Preprocessing
- Classification
- Regression
- Clustering
- Association Rules

Apart from these data mining concepts, it also provides the tools for Visualization. Weka has provided the package manager that can be used to install many learning schemes and tools. Package manager can be found in tools menu. It is an open source freely available tool issued under GNU general public license.

Weka tool provides the following application tools which can be used for various data mining stages using its inbuilt written algorithms as,

- Explorer
- Experimenter
- KnowledgeFlow
- Workbench
- Simple CLI

Weka has the mostly used application known as *Explorer*, which is the GUI based Integrated Development Environment (IDE) used to provide the system written algorithm for various data mining concepts such Preprocessing of data, classification, Clustering, Association Rules, Clustering etc.

CHAPTER 4
RESULTS AND DISCUSSION

In our research work, we have done the classification of a movie dataset using the subtitle of the movie trailer into various labels such as action, drama and romance. For our research work, we have used the data mining system tool WEKA 3.9 which has the system written algorithm for various phases of machine learning approach.

4.1 EXPERIMENTAL RESULTS

In our research work, the movie trailer data has been loaded in Weka during preprocessing phase which Weka refused to accept, as the data has been stored in text format. However, we have been able to load our data using system built converter, TextDirectoryLoader for the preprocessing. After applying the filter StringToWordVector, strings has been split into words as shown in the figure 4.1.

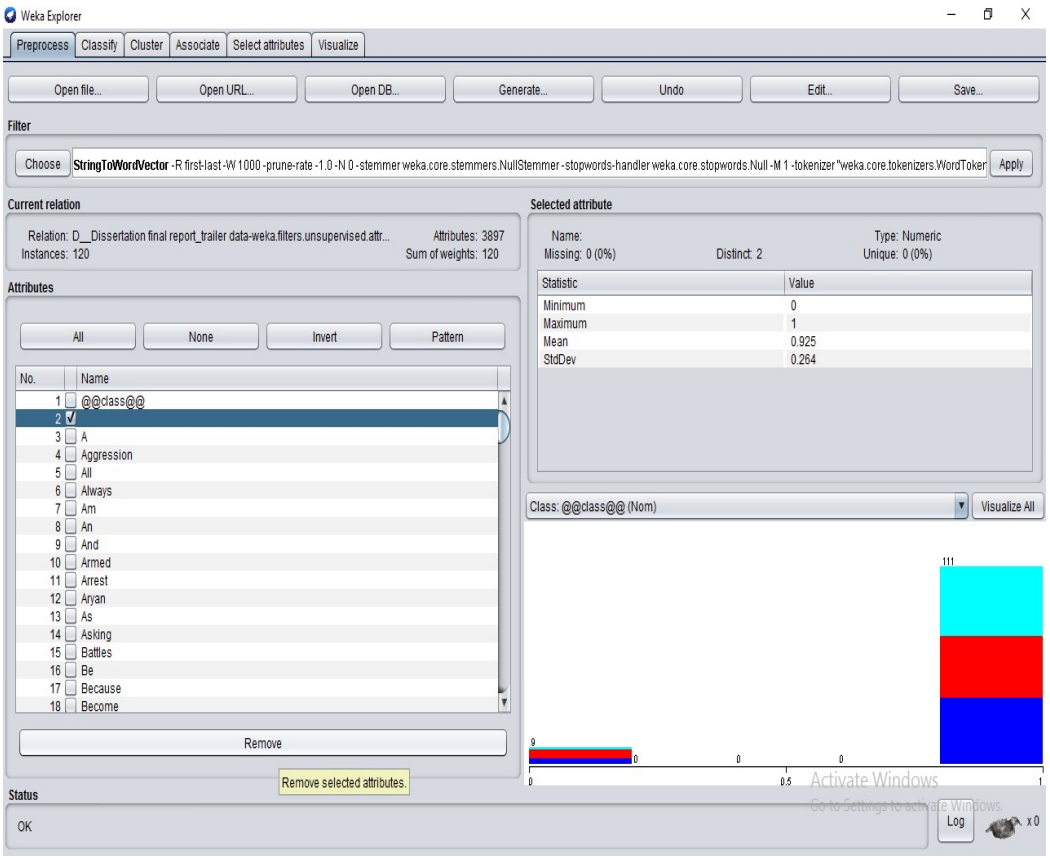

Figure 4.1: Applying StringToWordVector filter on movie dataset

After preprocessing of dataset, Support Vector Machine (SVM) Classifier has been applied for the classification of movie data into three labels such as action, drama and romance. Various test options have been applied to for better classification accuracy as shown in the figure 4.2.

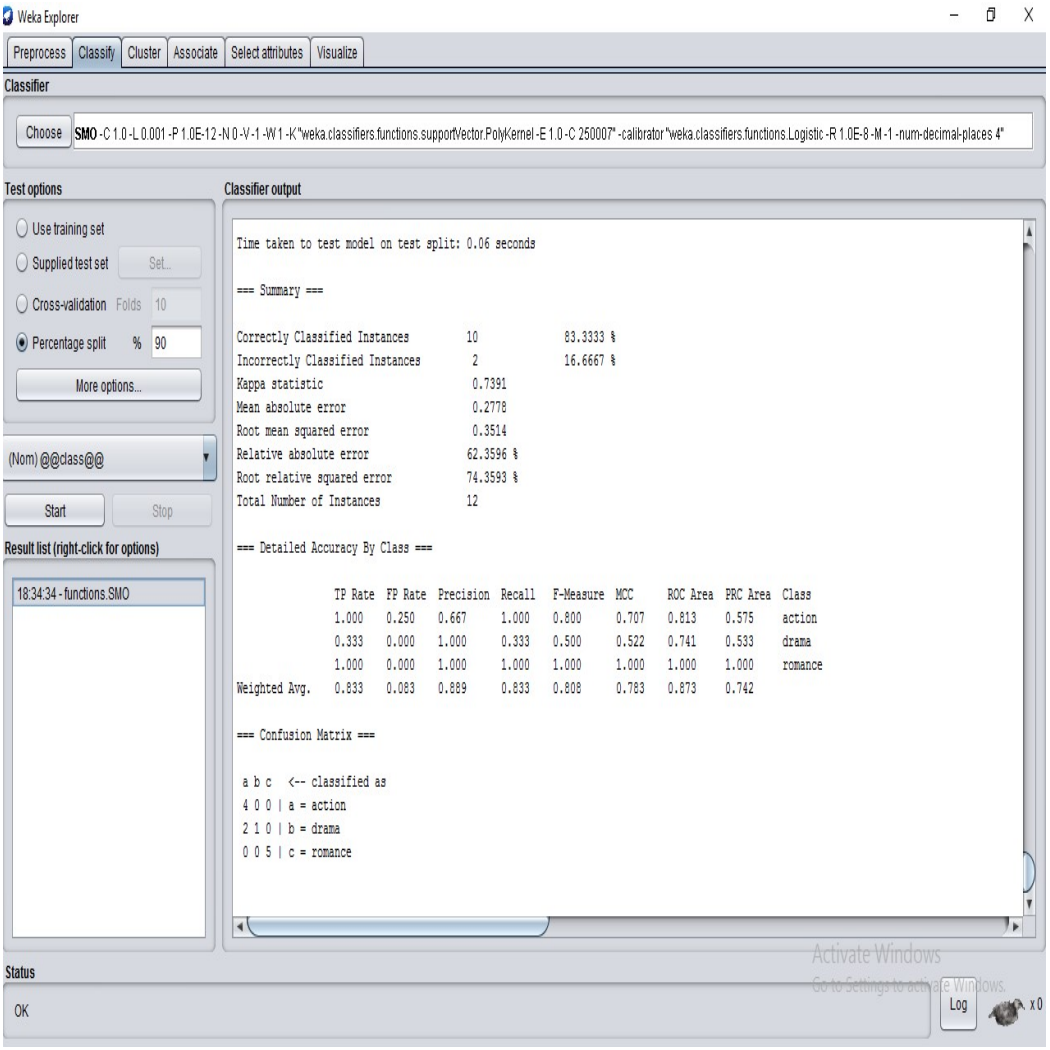

Figure 4.2: Applying SVM classifier for movie dataset

Support Vector Machine Classifier has been used to classify the preprocessed movie dataset, using Percentage Split and Cross Validation testing methods. The result provided that when *Percentage Split* has been used as testing option, there has been best classification accuracy achieved is 83.33%. Table 4.1 shows the classification accuracy achieved by SVM classifier with different value of percentage split for testing of data whereas figure 4.3 shows the bar graph for the variation in classification accuracy with different values of percentage split.

Table 4.1: SVM Classification Accuracy for Percentage Split

PERCENTAGE SPLIT	CLASSIFICATION ACCURACY
35%	39.74%
45%	48.48%
57%	53.85%
71%	65.71%
73%	62.50%
76%	62.07%
80%	54.17%
82%	59.09%
85%	66.67%
89%	69.23%
90%	83.33%

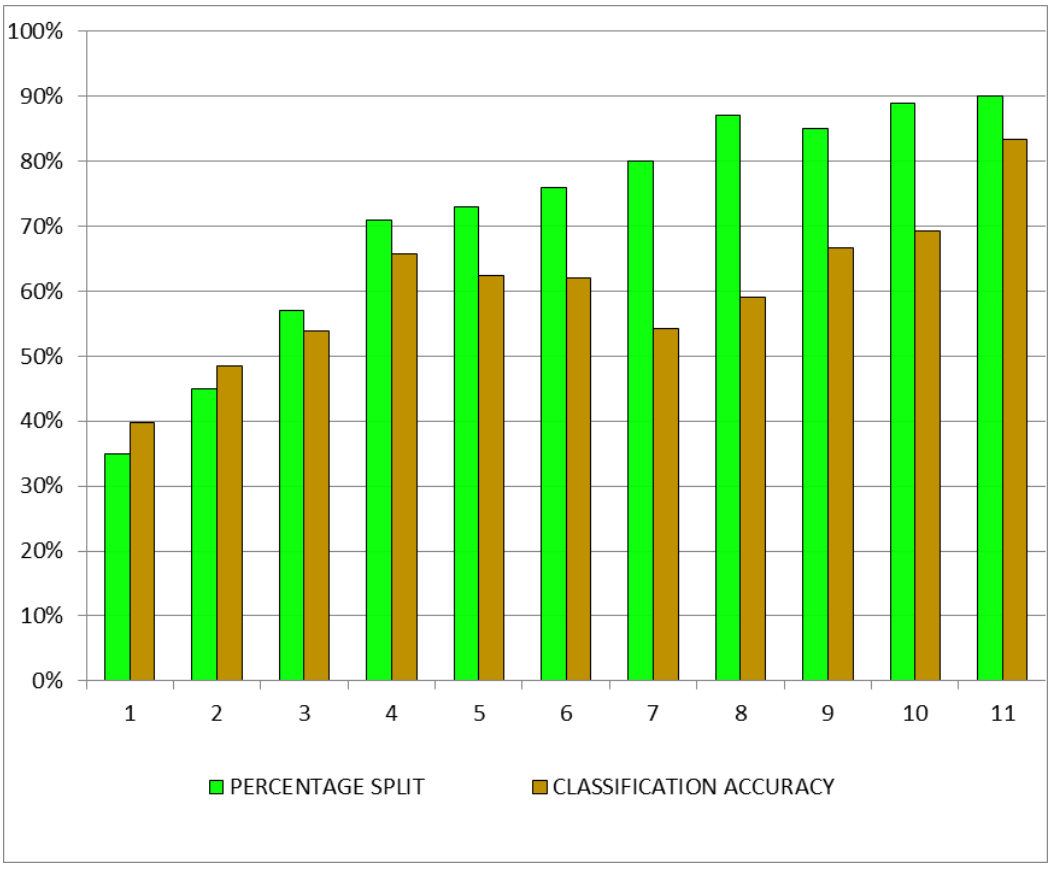

Fig. 4.3: Bar Graph of SVM Classification Accuracy for Percentage Split

SVM classification accuracy for different values of folds for Cross Validation testing method can be shown in table 4.2 and by the Bar Graph in the figure 4.4.

Table 4.2: SVM Classification Accuracy for Cross Validation

CROSS VALIDATION FOLDS	CLASSIFICATION ACCURACY (%)
5	58.33
10	60
15	55
20	58.33
25	55.83
30	60
50	55.83

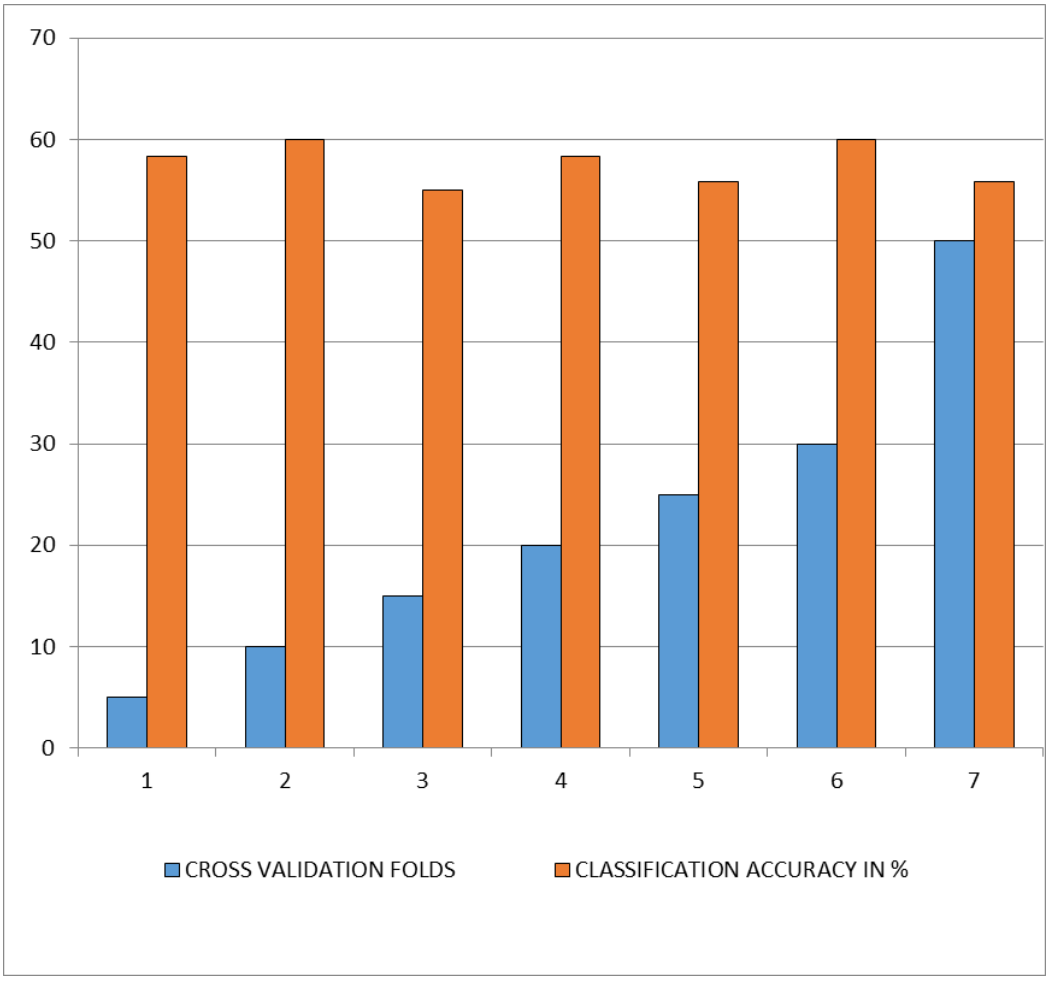

Fig. 4.4: Bar Graph of SVM classification accuracy for Cross Validation

We have got the best Classification Accuracy of 83.33% for SVM Classifier using Percentage Split testing option. So Percentage Split testing option has been best for our movie dataset. Figure 4.5 shows the detail result for SVM classifier using 90% percentage split (90% training data and 10% testing data).

```
Time taken to test model on test split: 0.04 seconds

=== Summary ===

Correctly Classified Instances          10               83.3333 %
Incorrectly Classified Instances         2               16.6667 %
Kappa statistic                          0.7391
Mean absolute error                      0.2778
Root mean squared error                  0.3514
Relative absolute error                 62.3596 %
Root relative squared error             74.3593 %
Total Number of Instances               12

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.250    0.667      1.000    0.800      0.707    0.813     0.575     action
               0.333    0.000    1.000      0.333    0.500      0.522    0.741     0.533     drama
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     romance
Weighted Avg.  0.833    0.083    0.889      0.833    0.808      0.783    0.873     0.742

=== Confusion Matrix ===

 a b c   <-- classified as
 4 0 0 | a = action
 2 1 0 | b = drama
 0 0 5 | c = romance
```

Figure 4.5: Result of SVM Classifier on 90% Percentage Split

Performance of the Classifier on a test dataset for which the true values are known is shown through a table called Confusion Matrix. For Support Vector Machine (SVM) classifier, the confusion matrix has been shown in the Table 4.3 which shows the performance of the Classifier for 90% Percentage Split testing option. Figure 4.6 shows the confusion matrix for SVM classifier using 90% of training dataset to provide the best accuracy for the movie dataset using Bar Graph.

Table 4.3: Confusion matrix for SVM using 90% Percentage Split

	a	B	C
Action=a	4	0	0
Drama=b	2	1	0
Romance=c	0	0	5

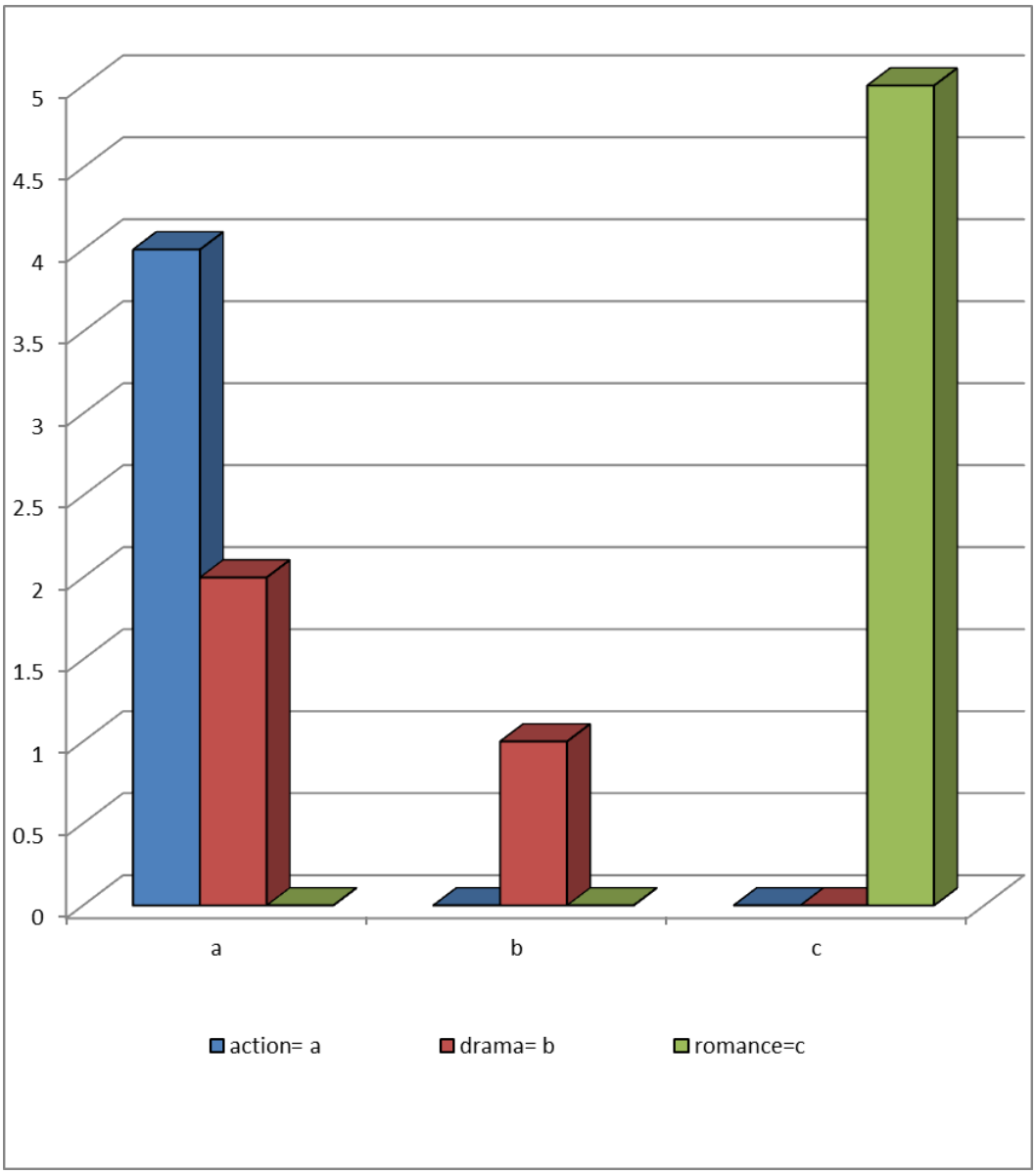

Figure 4.6: Bar Graph of Confusion matrix for SVM using 90% Percentage Split

Following are the basic terms which define the Confusion Matrix:

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

Apart from these basic terms following are the terms used to determine the accuracy of the classification as shown in the table 4.4, for the SVM Classifier using 90% Percentage Split.

Table 4.4: Various terms used to determine the accuracy of the classification

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC area	PRC area	Class
	1.00	0.25	0.667	1.000	0.800	0.707	0.81	0.58	Action
	0.33	0.00	1.000	0.333	0.500	0.522	0.74	0.53	Drama
	1.00	0.00	1.000	1.000	1.000	1.000	1.00	1.00	romance
Wt. Avg	0.83	0.08	0.889	0.833	0.808	0.783	0.87	0.74	

In our research work, these values can be manually calculated of 90% Percentage Split for the SVM Classifier with the help of following mathematical formulas:

- *Accuracy*

  Correctly classified instances accuracy= (True values/Total values)

  Correctly classified instances accuracy= 10/12= 0.833= 83.33%

  Incorrectly classified instances accuracy= (False values/Total values)

  Incorrectly classified instances accuracy= 2/12= 0.166= 16.66%

- *TP (True Positive) rate*

  For Class action,

  TP= 4/4= 1.00

  For Class drama,

  TP= 1/3= 0.33

For Class romance,

TP= 5/5= 1.00

Similarly, all the values for each term can be manually calculated using mathematical formula to match with the obtained values through experiment. Figure 4.7 showing the values of various terms used to determine the classification accuracy.

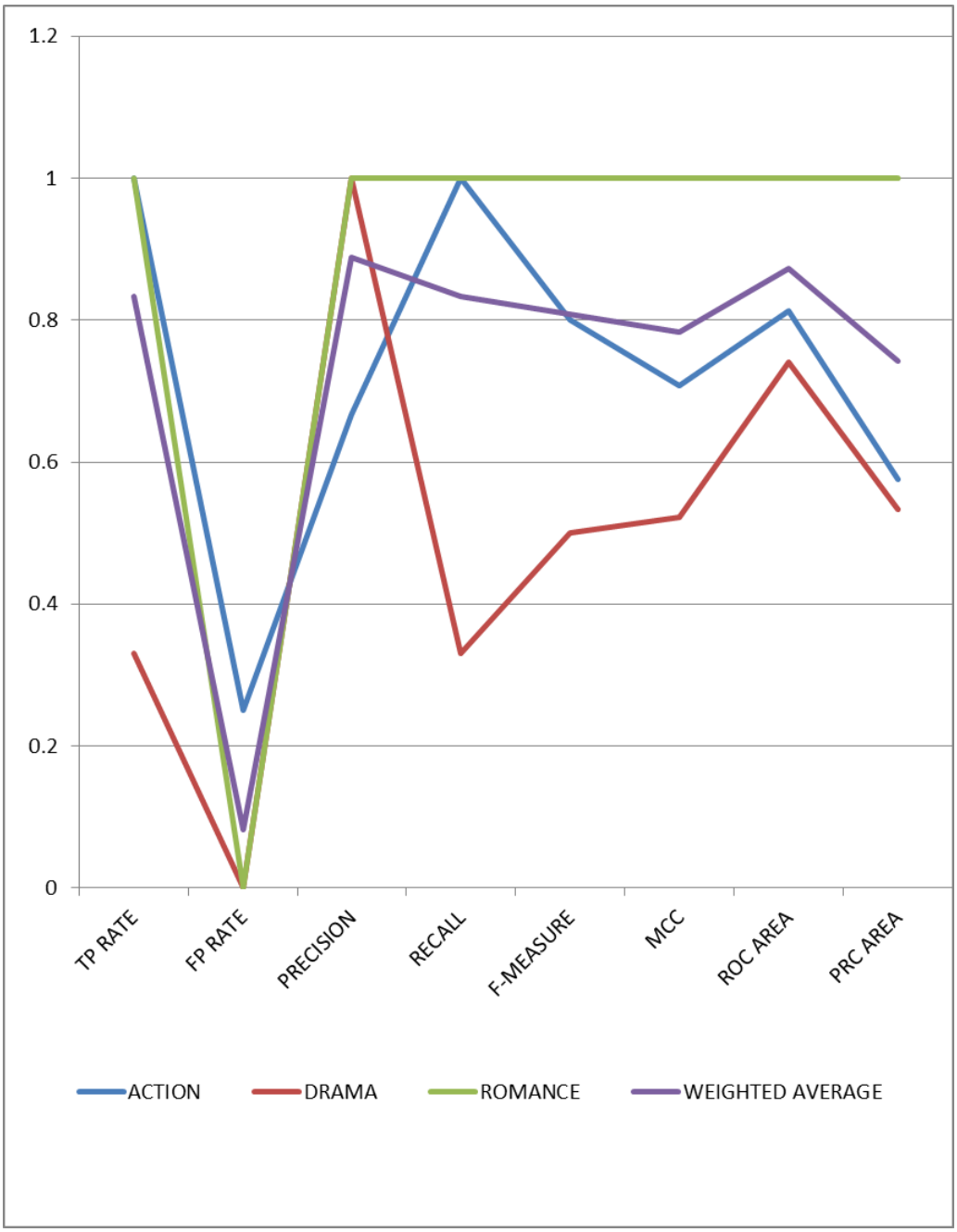

Figure 4.7: Line graph showing values of various terms used to determine accuracy

Weka has provided the Visualization tool to showcase the result using various visualization measures such as Cost/Benefit analysis, threshold curve, margin curve and cost curve. Figure 4.8 shows the Cost/Benefit analysis graph of the action class for SVM Classifier, testing with 90% of training dataset.

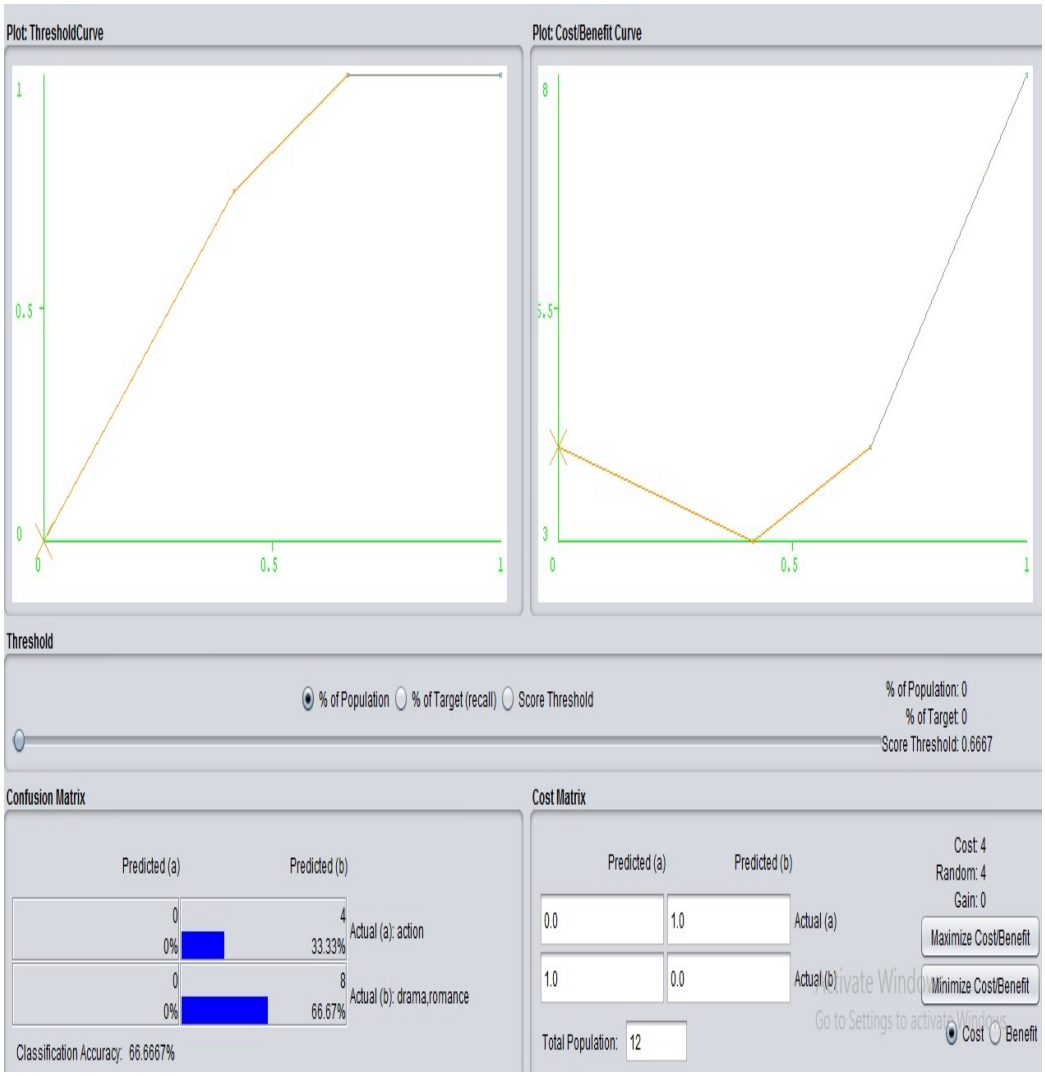

Figure 4.8: Cost/Benefit analysis graph of action class for SVM Classifier

Cost/Benefit analysis graph has shown the plot of Threshold curve and Cost/Benefit curve for each class such as action, drama and romance. It has also presented the Confusion matrix and Cost matrix for each class. Figure 4.9 shows the Cost/Benefit analysis graph for class *drama* whereas figure 4.10 shows the Cost/Benefit analysis graph for class *romance*. Confusion matrix shown in the graph has provided the classification accuracy of each class. For example, Classification accuracy for class drama has been shown as 75% for the SVM classifier using Percentage Split method.

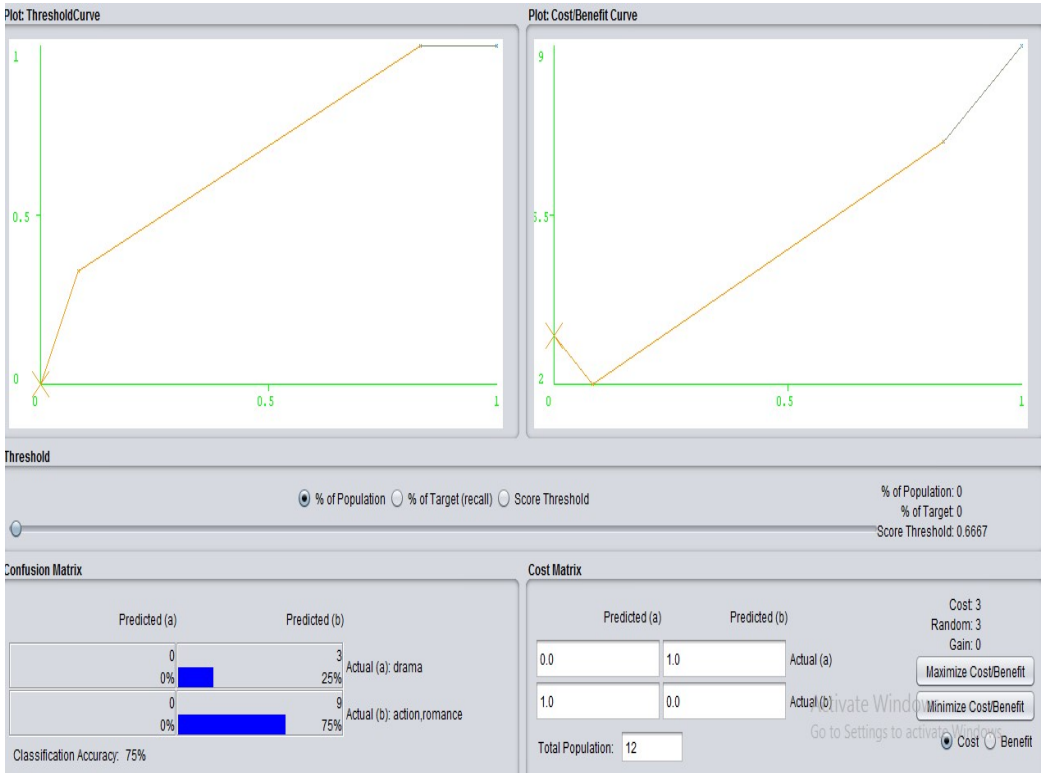

Figure 4.9: Cost/Benefit analysis graph of drama class for SVM Classifier

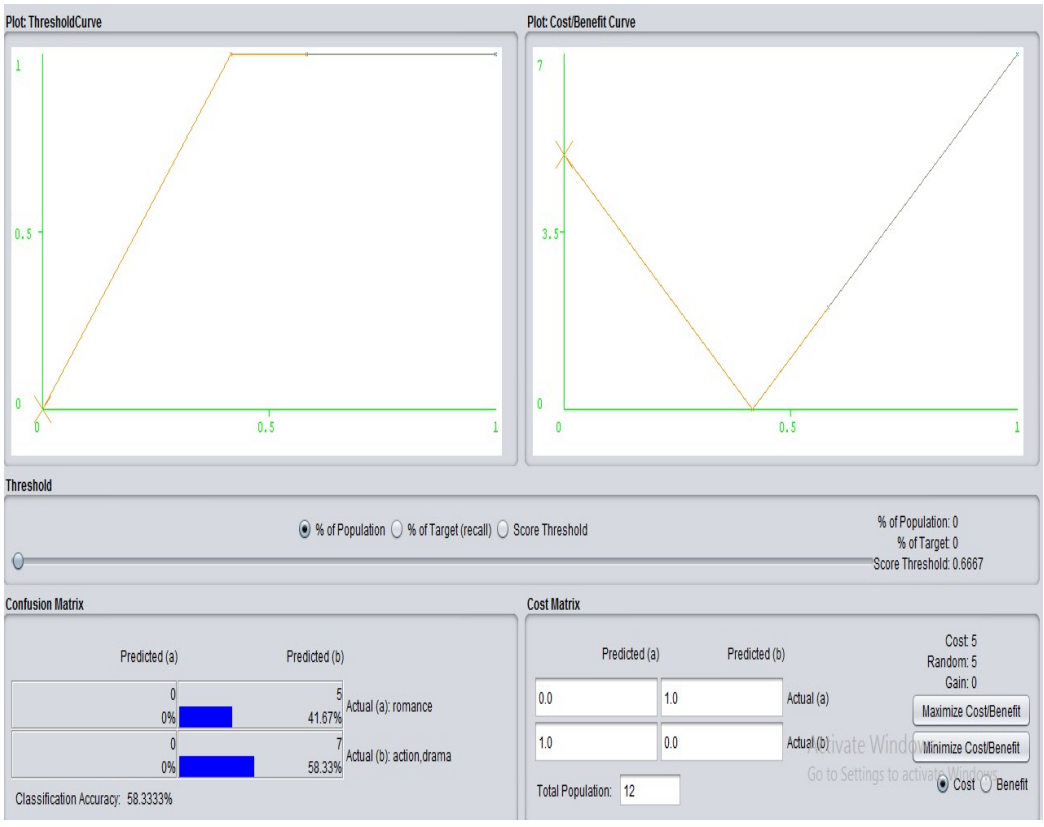

Figure 4.10: Cost/Benefit analysis graph of romance class for SVM Classifier

4.2 COMPARISON OF SVM WITH RANDOM FOREST

In our research work, we have used another Classifier Random Forest for the compare based analysis with SVM classifier, to find out the Classifier with best classification accuracy and performance. Performance of a Classifier has been shown with the help of Confusion Matrix.

Table 4.5 shows a comparison of Classification Accuracy result between the SVM and Random Forest (RF) classifier for same values of Percentage Split testing options.

Table 4.5: A Comparison of Classification Accuracy between SVM and RF for percentage split

PERCENTAGE SPLIT	CLASSIFICATION ACCURACY FOR SVM	CLASSIFICATION ACCURACY FOR RF
35%	39.74%	39.74%
45%	48.48%	42.42%
57%	53.85%	51.92%
71%	65.71%	57.14%
73%	62.50%	59.38%
76%	62.07%	51.72%
80%	54.17%	58.33%
82%	59.09%	50%
85%	66.67%	44.44%
89%	69.23%	61.54%
90%	83.33%	75%

A comparison of Classification accuracy result between Support Vector Machine (SVM) and Random Forest (RF) for the percentage split testing option has been shown in the figure 4.11 using Bar Graph.

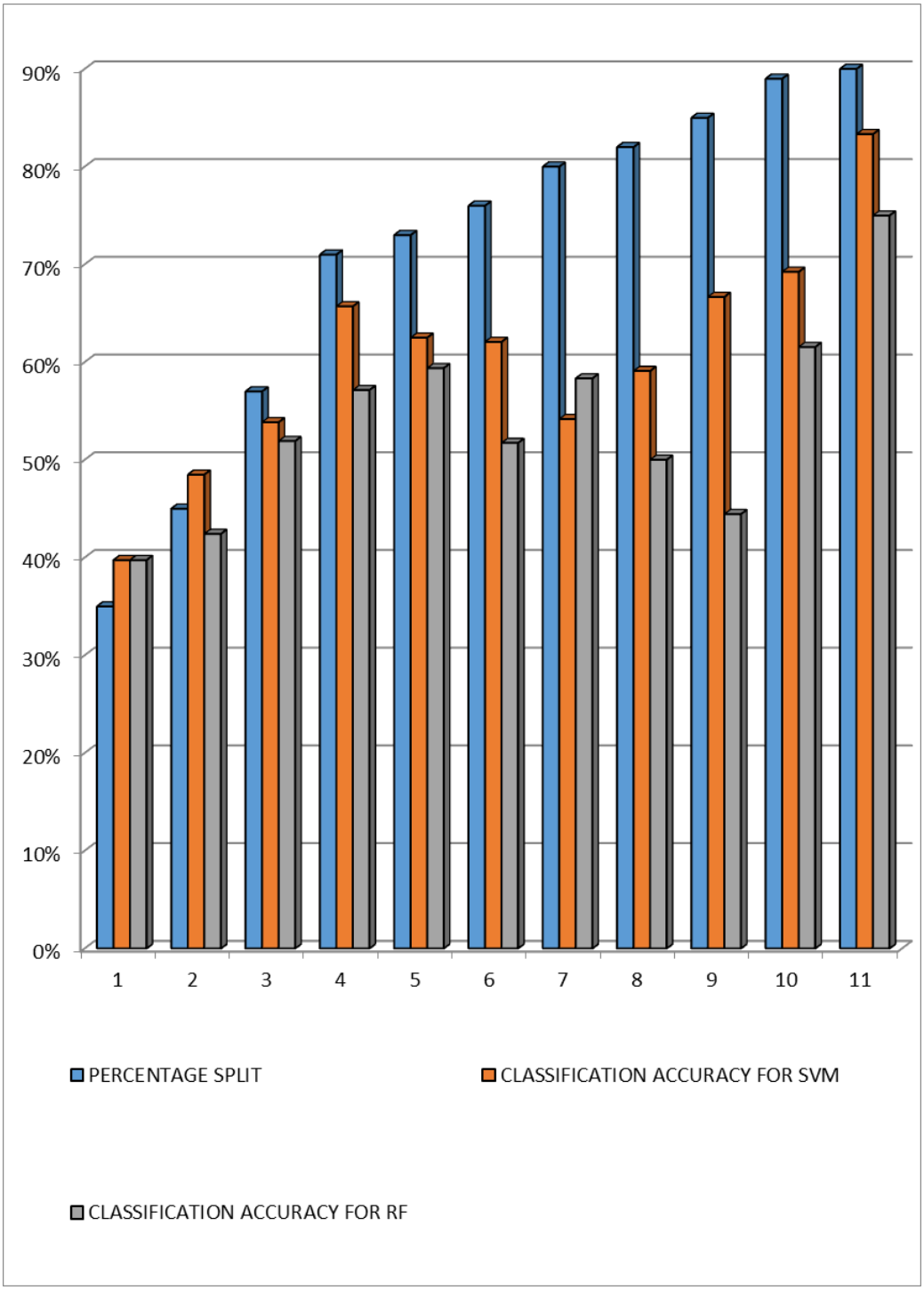

Figure 4.11: Bar Graph showing comparison of Classification accuracy between SVM and RF for percentage split

Similarly a comparison for Classification Accuracy between SVM and RF has been drawn for Cross Validation and shown through the table 4.6. Figure 4.12 has been used to show the same comparison through Bar Graph.

Table 4.6: A Comparison for classification accuracy between SVM and RF for Cross Validation

CROSS VALIDATION FOLDS	CLASSIFICATION ACCURACY FOR SVM (IN %)	CLASSIFICATION ACCURACY FOR RF (IN %)
5	58.33	46.67
10	60	51.67
15	55	45
20	58.33	51.67
25	55.83	50.83
30	60	50.83
50	55.83	50.83

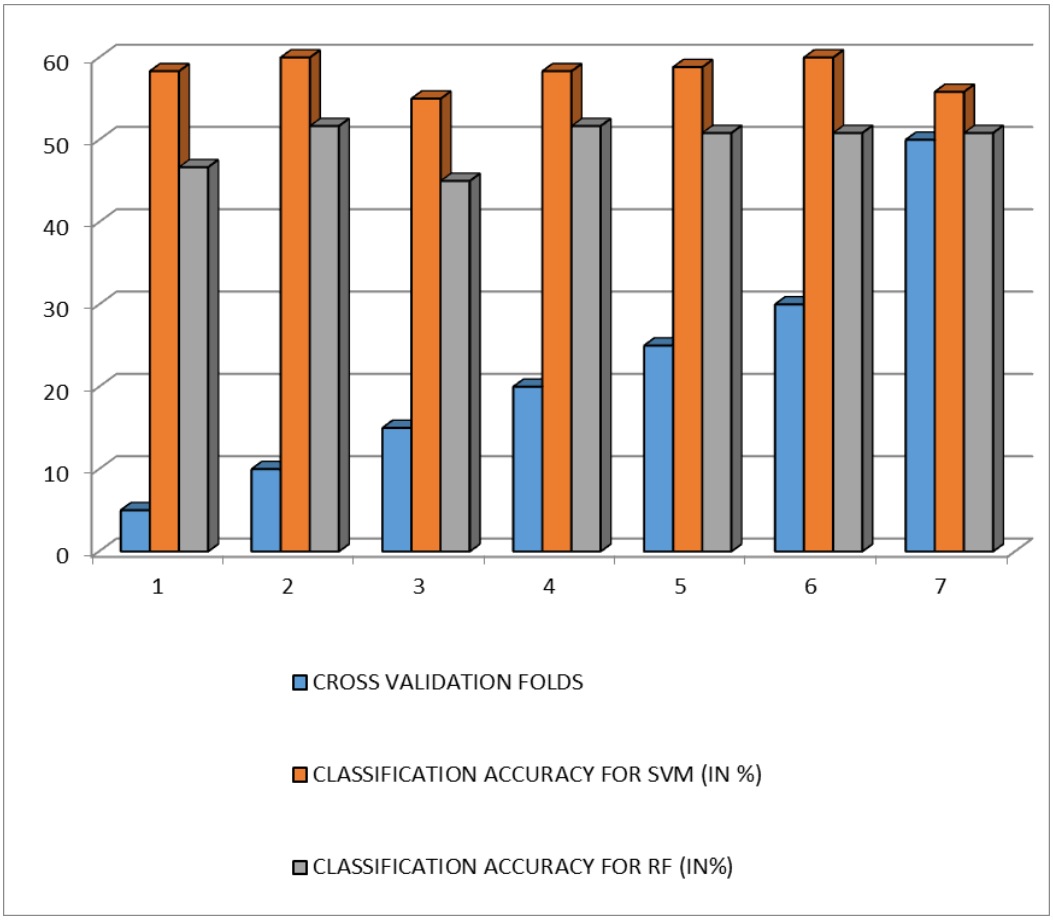

Figure 4.12: Bar Graph showing Comparison of Classification Accuracy between SVM and RF for Cross Validation

From table 4.5, it has been observed that Support Vector Machine (SVM) has the best classification accuracy of 83.33% in comparison of Random Forest (RF) with 75% best classification accuracy for Percentage Split testing option.

From table 4.6, it has been observed that SVM has the best classification accuracy of 60% in comparison of Random Forest with 51.67% best classification accuracy for cross validation testing option.

From above statements, it has been observed that Support Vector Machine (SVM) Classifier produced best Classification Accuracy for our testing movie dataset using both testing options.

Confusion Matrix has been used to describe the performance of a classifier using a table having TP and TN values. In our research work, we have compared the confusion matrix of Support Vector Machine and Random Forest to find out the Classifier who has generated the best performance for our movie trailer dataset. Table 4.7 shows the confusion matrix for Random Forest Classifier using 90% Percentage Split testing option.

Table 4.7: Confusion Matrix for RF using 90% Percentage Split

	a	**b**	**C**
Action=a	3	1	0
Drama=b	1	2	0
Romance=c	1	0	4

From table 4.3 and 4.7, it has been observed that Support Vector Machine (SVM) Classifier has given better performance than Random Forest (RF) Classifier. Confusion Matrix has been discussed in table 4.3 for SVM Classifier whereas Confusion Matrix of Random Forest has been discussed in the table 4.7.

We have also showed the comparison between SVM Classifier and Random Forest Classifier through Bar graph. Bar Graph showing the Confusion Matrix of Random Forest Classifier has been shown in the figure 4.13. Figure 4.6 and Figure 4.13 has showed the comparison of Confusion Matrix between SVM and Random Forest through Bar Graph.

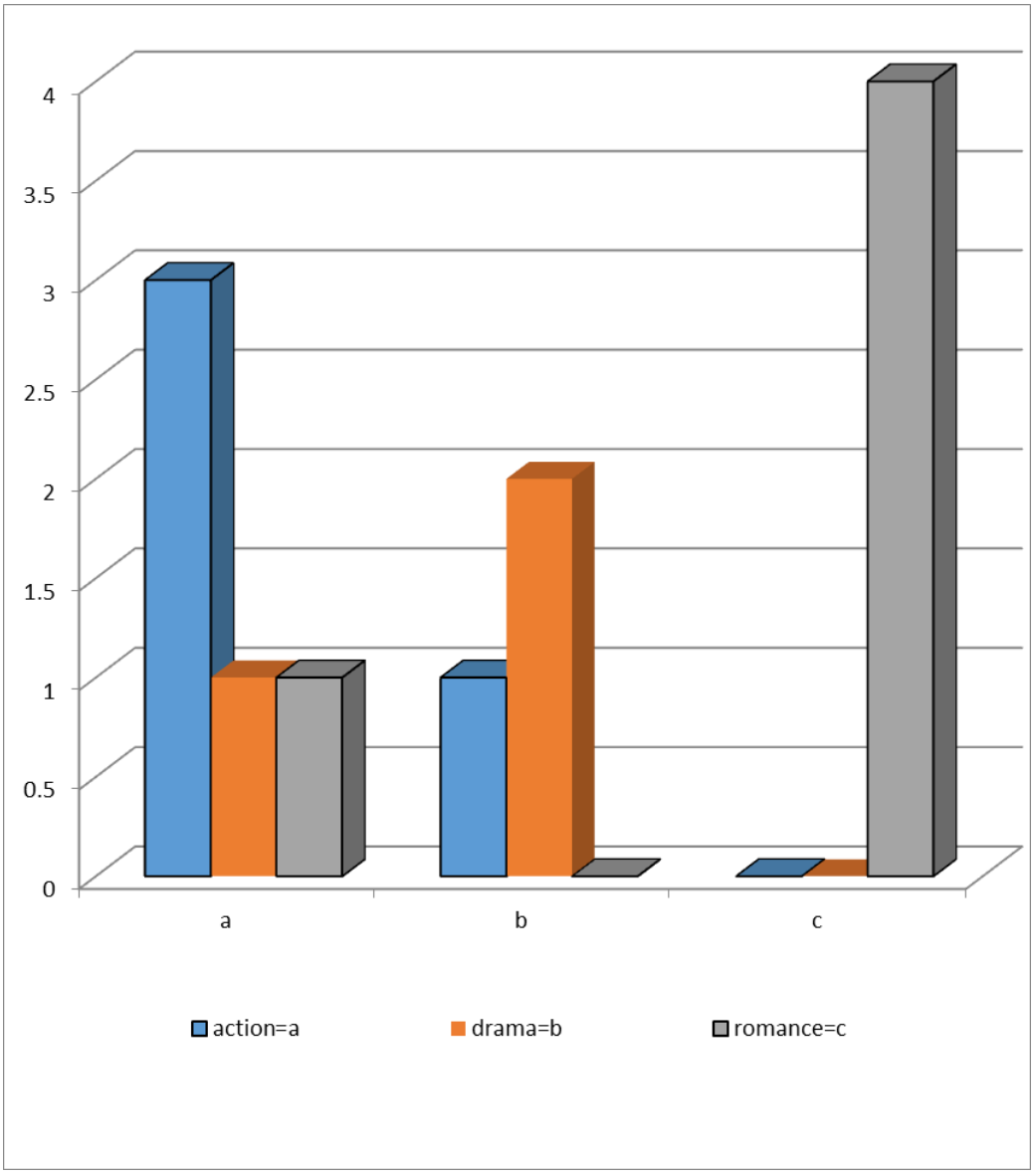

Figure 4.13: Bar Graph showing the Confusion Matrix for RF using 90% Percentage Split

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

In our research work, we have used the Support Vector Machine (SVM) Classifier to classify the movie trailer dataset into three labels i.e. action, drama and romance. Confusion Matrix has shown the performance of the Classifiers. Classification Accuracy has been recorded using Percentage Split and Cross Validation testing methods.

We have used Random Forest classifier to compare the *Classification Accuracy* and *Performance* with SVM. By analyzing the results of both the classifiers, we have concluded that Support Vector Machine (SVM) has been the better classifier, providing the best accuracy of 83.33% compare to Random Forest with best accuracy of 75%.

In our research work, we have faced difficulties in extracting data in transcript or subtitle form from the movie trailers. We have observed that the year before 2012, most of the movie trailers have been released without the subtitle or transcript. Another difficulty we have faced in the form of movie trailers having subtitles in languages other than English.

Testing has been done for the extracted dataset which has been stored in text document form. Weka tool has been splitting the whole dataset randomly using various testing options. One difficulty with this type of analysis has been that we cannot classify the specified movie trailer into respective genre class.

5.2 FUTURE SCOPE

We can extend this project as it can classify the specified movie trailer into their respective genre class. We can generate an interface which takes input for specified movie and after analyzing the data with existing database can predict the class.

We can generate the text file using the audio of the movie trailer for those which do not have the subtitle for their movie trailer and used that text document for classification of the movie trailer in respective genre class.

LIST OF REFERENCES

[1] A. Abbasi, A. Hassan, and M. Dhar, "Benchmarking Twitter Sentiment Analysis Tools," *Proc. Ninth Int. Conf. Lang. Resour. Eval.*, pp. 823–829, 2014.

[2] C. Cai and L. Li, "New Words Enlightened Sentiment Analysis in Social Media," pp. 202–204, 2016.

[3] L. F. S. Coletta, N. F. F. De Silva, E. R. Hruschka, and E. R. Hruschka, "Combining classification and clustering for tweet sentiment analysis," *Proc. - 2014 Brazilian Conf. Intell. Syst. BRACIS 2014*, pp. 210–215, 2014.

[4] A. D'Andrea, F. Ferri, P. Grifoni, and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation," *Int. J. Comput. Appl.*, vol. 125, no. 3, pp. 26–33, 2015.

[5] D. V. N. Devi, "Sentiment Analysis Using Harn Algorithm," 2016.

[6] Rushlene K. Bakshi, Navneet Kaur, Ravneet Kaur, Gurpreet Kaur "Opinion Mining and Sentiment Analysis," pp. 452–455, 2016.

[7] D. Fahey, "AffinityFinder : A System for Deriving Hidden Affinity Relationships on Twitter Utilizing Sentiment Analysis," pp. 4–7, 2016.

[8] M. Furini and M. Montangero, "TSentiment: On gamifying Twitter sentiment analysis," *2016 IEEE Symp. Comput. Commun.*, pp. 91–96, 2016.

[9] A. Gupte, S. Joshi, P. Gadgul, and A. Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis," *...) Int. J.* vol. 5, no. 5, pp. 6261–6264, 2014.

[10] F. Luo, C. Li and Z. Cao,"Affective-feature-based sentiment analysis using SVM classifier," *2016 IEEE 20th Int. Conf. Comput. Support. Coop. Work Des.*, pp. 276–281, 2016.

[11] B. Liu, “Sentiment analysis: A multifaceted problem,” *IEEE Intell. Syst.*, vol. 25, no. 3, pp. 76–80, 2010.

[12] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[13] R. Mihalcea and L. Morency, “Mulitmodal Sentiment Analysis of Spanish Online Videos,” vol. X, no. 3, 2011.

[14] S. S. Minab, “Online Analysis of Sentiment on Twitter,” no. Ictck, pp. 11–12, 2015.

[15] P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha, “An improved sentiment analysis of online movie reviews based on clustering for box-office prediction,” *Int. Conf. Comput. Commun. Autom. ICCCA 2015*, pp. 933–937, 2015.

[16] B. Pang and L. Lee, “A Sentimental Education: Sentiment Analysis using Subjectivity Summation based on Minimum Cuts,” *ACL ’04 Proc. 42nd Annu. Meet. Assoc. Comput. Linguist.*, p. 271, 2004.

[17] D. Y. Priyanka and R. Senthilkumar, “Sampling techniques for streaming dataset using sentiment analysis,” *2016 Int. Conf. Recent Trends Inf. Technol.*, pp. 1–6, 2016.

[18] Prachi Kasbekar, Gajanan Gaikwad, “ Multiclass Classification and class based Sentiment Analysis For Hindi Language,” no. April, pp. 512–518, 2013.

[19] M. Thelwall and R. Prabowo, “Sentiment analysis: A combined approach,” *J. Informetr.*, vol. 3, no. 2, pp. 143–157, 2009.

[20] P. Tripathi, S. K. Vishwakarma, and A. Lala, “Sentiment Analysis of English Tweets Using RapidMiner,” 2015.

[21] K. Kwon, Y. Jeon, C. Cho, and H. Park, “Sentiment trend analysis in social web environments,” pp. 261–268, 2017.

[22] P. Mishra, R. Rajnish, and P. Kumar, “Sentiment analysis of Twitter data: Case study on digital India,” *2016 Int. Conf. Inf. Technol. - Next Gener. IT Summit Theme - Internet Things Connect your Worlds*, pp. 148–153, 2016.

[23] T. P. Sahu and S. Ahuja, “Sentiment analysis of movie reviews: A study on feature selection & classification algorithms,” *2016 Int. Conf. Microelectron. Comput. Commun.*, pp. 1–6, 2016.

[24] P. Sharma, “Feature level Sentiment Analysis on Movie Reviews,” no. October, pp. 306–311, 2016.

[25] A. Goyal, “Sentiment Analysis for Movie Reviews,” pp. 1–8.

[26] S. Ezzat, N. El Gayar, and M. M. Ghanem,“Sentiment Analysis of Call Centre Audio Conversations using Text Classification,” *Int. J. Comput.,* vol. 4, pp. 619–627, 2012.

[27] X. Fang and J. Zhan, “Sentiment analysis using product review data,” *J. Big Data*, vol. 2, no. 1, p. 5, 2015.

[28] S. K. Tiwari, M. Kumar, and M. A. Alam, “A Literature Review on Opinion Mining and Sentiment Analysis,” vol. 5, no. 4, pp. 1–31, 2015.

[29] A. Goel, J. Gautam, and S. Kumar, “Real time sentiment analysis of tweets using Naive Bayes,” *2016 2nd Int. Conf. Next Gener. Comput. Technol.*, no. October, pp. 257–261, 2016.

[30] J. Zhao and X. Gui, “Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis,” *IEEE Access*, vol. 3536, no. c, pp. 1–1, 2017.

APPENDIX

NLP : Natural Language Processing

NWLB : New Word Lexicon-based

PCA : Principal Component Analysis

HSWN : Hindi Senti Word Net

ML : Machine Learning

RF : Random Forest

SVM : Support Vector Machine

TP : True Positive

TN : True Negative

FP : False Positive

FN : False Negative

RDF : Resource Description Framework

Sk-Learn : Scikit-Learn

KNN : K-Nearest Neighbors

NLTK : Natural Language Tool Kit

ACO : Ant Colony Optimization

API : Application Program Interface