**Sentiment Analysis of Movie Reviews**

****

Sikandar Ali

CIIT/FA14-BCS-095/VHR

Abida Zaman

CIIT/SP14-BCS-061/VHR

BSCS Thesis

COMSATS University Islamabad,

Vehari campus , Pakistan

SP-2018



**COMSATS University Islamabad , Vehari campus.**

**Sentiment analysis of Movie Review**

A Thesis Presented to

COMSATS University Islamabad , Vehari campus

In partial fulfillment

Of the requirements for the degree of

Bachelor of Science

in

Computer Science

Department of Computer Science

By

Sikandar Ali

CIIT/FA14-BCS-095/VHR

Abida Zaman

CIIT/SP14-BCS-061/VHR

SP,18

**Sentiment analysis of Movie Review**

An undergraduate thesis submitted to the Department of Computer Science as partial fulfillment of the requirements for the award of Degree of Bachelor of Science in Computer Science

|  |  |
| --- | --- |
| Name | Registration Number |
| Sikandar Ali | CIIT/FA14-BCS-095/VHR |
| Abida Zaman | CIIT/SP14-BCS-061/VHR |

**Supervisor**

Mr. Zahid Anwer

Department of Computer Science

Signature of Supervisor: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_

COMSATS University Islamabad, Vehari campus

June, 2018

Final Approval

This Thesis Titled

**Sentiment Analysis of Movie Review**

By

*Sikandar Ali*

*CIIT/FA14-BCS-095/VHR*

*Abida Zaman*

*CIIT/SP14-BCS-061/VHR*

External Examiner: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr. \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

Designation. \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

Supervisor: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Mr. Zahid Anwar Research Associate, Department of Computer Science COMSATS Vehari campus

Internal Examiner: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr./Mr. \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

Designation:\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

Department of Computer Science COMSATS

University Islamabad, Vehari campus

Convener Project Committee: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Dr./Mr. \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

Designation:\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

Department of Computer Science COMSATS

University Islamabad, Vehari campus

**DECLARATION**

It is certified that the research work presented in this thesis is to the best of my knowledge my own. All sources used and any help received in the preparation of this dissertation have been acknowledged. Hereby, it is declared that this material is not submitted, either in whole or in part, for any other degree at this or any other institution.

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Signature of the Student: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Sikandar Ali

CIIT/FA14-BCS-095/VHR

Signature of the Student: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Abida Zaman

CIIT/SP14-BCS-061/VHR

**CERTIFICATE**

It is certified that Mr. Sikandar Ali (CIIT/--/VHR) and Abida Zaman (CIIT/--/VHR) have carried out all the work related to this thesis under my supervision at the Department of Computer Science COMSATS Institute of Information Technology, Vehari. And the work fulfills the requirements for award of BS degree.

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_ Supervisor: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Mr.Zahid Anwer

Lecturer, Department of Computer Science

CUI, Vehari

**DEDICATION**

**To Almighty ALLAH and the Holy Prophet Muhammad (P.B.U.H)**

**&**

**My Loving Family**

**ACKNOWLEDGMENTS**

First and foremost, we would like to thank Head of Department computer science department of CIIT Vehari (Dr. Rafiq Mufti ) and our project supervisor (Sir Zahid Anwar) for supporting and encouraging us all the time giving us opportunity to enhance our technical skills and complete this project. We would like to thank our parents and family members for supporting us during our academic career, who always support us so that we could focus on our project work. Finally, we would like to thank all our class mates and friends whose names are not mentioned here but have helped us in some way to accomplish our project work.

**Abstract**

During the last years, sentiment analysis has become a very popular task in Natural language processing. Many people use the internet for the seeking of opinion. With the increasing amount of user-based content on the we there has been emergence of research fields that uses sentiment analysis to a advantage of and process this data. This could result more satisfied customers as more relevant information will be easier to be find. Affecting analysis of text is usually obtainable as the problem of automatically identifying a representative sentimental category or scoring the text within a set of sentiment dimensions. However, most existing approaches determine these categories and dimensions by matching the terms in the text with those presented in an affective lexicon. We present application system that uses movie review from the user’s movie account to show the affective sentiment of movie review. We use Multinomial Naïve Bayes algorithm and Support Vector Machine(SVM) for train our data , and accuracy is 96% and 85% in it and use a dataset of 2000 movie reviews from corpus for getting the accuracy. Then we do testing on a large dataset of 25000 movie reviews to get the sentiments.

**Table Of Contents:**

**Chapter 1: Introduction**

1. Introduction: …………....................................................................................................14

1.1. Problem Definition……………........................................................................………14

1.2. Motivation and Context ................................................................................................15

1.3. Objectives......................................................................................................................15

1.4. Project Scope...…..............................................................................................………16

1.4.1. Dataset.....................................................................................................................16

1.4.2. Classification Algorithms........................................................................................16

1.5. Tools /Technology……………………………………………………………….……17

1.5.1. Tools………………………………………............................................................17

1.5.2. Language………………………………………………………………………….17

1.5.3. Operating System...……………...……………………….……………………….17

1.6 Approach………………………………………………………………………………18

1.7 Main Contribution…………………………………………………………………..…18

1.8. Structure of Report……………………………………………………………….…...18

**Chapter 2: Literature Review**

2.1. Sentiment Analysis and Opinion Mining……………………………………………...20

2.2 Online Paper Review Analysis………………………………….……………………...20

2.3.Sentiment Analysis of Review Datasets using Naïve Bayes…………………………...21

**Chapter 3: Background and Related Work**

3 Background and Related Work .....................................................................................…..23

3.1 Sentiment Analysis...........................................................................................................23

3.2 Basic Concepts .................................................................................................................23

3.2.1. Sentiment feeling.............................................................................…….................23

3.2.2. Automated sentiment analysis..........................................................……………... 24

3.2.3 Emoticons...................................................................................................………...24

3.2.4 Features .....................................................................................................................24

3.3 Related Work sentiment Analysis..........................................................................…….26

**Chapter 4: System Overview**

4. System Overview ............................................................................................................28

4.1 Proposed System ...........................................................................................................28

4.2 Sentiment Recoginition..................................................................................................29

4.3 Data cleaning……………..............................................................................................29

4.4 Non- Functional Requirements......................................................................................30

**Chapter 5: Experimental Setup**

5. Experimental Setup……………………………………………………………………...32

5.1 System Flow....................................................................................................................32

5.1.1 Obtaining the Datatset...............................................................................................33

5.1.1.2 Scikit-learn…………………………………………………………………….33

5.1.2 Preprocessing of data.................................................................................................33

5.1.2.1 StopWords............................................................................................................34

5.2 Feature Selection.............................................................................................................34

5.2.1 Count-Vectorizer.......................................................................................................34

5.2.2 TIDF Transformer.....................................................................................................35

5.3 Algorithms.......................................................................................................................36

5.3.1 Multinomial NB()……………………………………………………………..........36

5.3.2 Support Vector Machine…………………………………………………..………..37

5.4 Prediction……………………………………………………………………………….38

5.4.1 Numpy………………………………………………………………………………38

**Chapter # 6:Results and Discussion**

6. Results and Discussion…………………………………………………………….……..41

6.1. Evaluation Matrices…………………………………………………………………….41

6.2. Result by Multinomial NB() on the smallest Dataset…………………………………..44

6.3 Result by Support Vector Machine NB() on the smallest Dataset………………………45

6.4 Result by Multinomial NB() on the largest Dataset……………………………………..46

6.5Result by Support Vector Machine NB() on the largest Dataset…………………………47

6.6 Accuracy of small and Large Dataset……………………………………………………48

**Chapter # 7:Conclusion and Future Work**

7. Conclusion and Future Work…………………………………………………………..….50

**References**……………………………………………………………….....51

**Chapter # 1**

**Introduction**

**1.Introduction:**

Sentiment analysis, also called opinion mining, is the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes and emotions towards entities such as products, services, organizations individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g.…sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc.

Although social media is growing rapidly, still there is a need to get useful information from the data scattered on social media. All the data available on social media or social networking sites is raw, unstructured and spread all over. This research is based on sentiment analysis of movie review data. movie review, a microblogging site has emerged as a popular social media site that allows its users to send and read the posts of up to 140 characters known as “movie review”. The main issues in extracting sentiments from movie review are: abbreviations, lack of capitals, poor grammar, poor punctuations, and poor spellings movie review is selected for this research due to its limited amount of characters in which users can precisely describe their thoughts thereby handling the associated issues [1].

* 1. **Problem Definition :**

We address the task of determining sentiments expressed in text at the sentence level and assign automatically a score label to each word in the given text. The possible labels are positive, negative. Those are the two basic sentiment categories identified by Ekman (1992).The scope of this work is limited to determining the sentiments or opinions of sentences, as much as is evident from the written text. The criterion for determining what constitutes emotional text and what is the type of sentiments expressed is human judgment. It is common experience that just as opinions can be expressed in many ways, the same expression may be interpreted differently by different readers. The techniques introduced in this work were evaluated against data whose emotional orientation was judged the same by at least two humans. The scope of work is to label its sentiment negative, positive. We‘ll perform sentence level sentiment analysis[2].

**1.2. Motivation and Context**

The motivation for this work has come from the recent growing interest in the data sciences field. The rapid growth of the World Wide Web has facilitated increased online communication and opened up newer avenues for the general public to post their opinions online. This has led to generation of large amounts of online content rich in user opinions, sentiments, emotions, and evaluations. We need computational approaches to successfully analyze this online content, recognize and aggregate relevant information, and draw useful conclusions. Much of the current work in this direction has typically focused on recognizing the polarity of sentiment (positive/negative). Among the less explored data sciences areas is the recognition of types of sentiments and their intensity – the focus of this work.

Recognizing emotions conveyed by a text can provide an insight into the author’s intent and sentiment, and can lead to better understanding of the text’s content. The inspiration for this work has also come from studies in psychology, which focus on analyzing texts to gain a deeper understanding of the way people express different kinds of opinions. These studies are typically carried out in controlled laboratory settings, or drawn from academic writings or medical domain. Companies or governments are unable to take opinion of people manually Internet is big source they can use to find out what people think about their products and polices[3].

**Objective :**

Most subscription or e-commerce systems, contain a form of review system, but they are not always exploited to the fullest. Meaning the comments are not extracted, summarized or taken advantage of, beside showing comments that people can read. A limitation of such a system is that a user most likely don’t read more than a couple of the first comments, resulting in a user missing out on relevant information. For example, we are looking for a hotel to spend the night. The tenth comment, that we never read ,told that the beds were really hard. Since we never read that comment, we ordered the hotel and got disappointed of the beds. If the recommender system uses all the comments of a movie or a product, the user may experience better recommendations and more satisfaction in their own choices of products.

The scope of this thesis is limited to testing out theories about how we can extract opinions from natural language text. The scope do not include developing a recommender system. Our idea is that adjectives often determines the sentiment of a sentence. The main research goal of this thesis is to investigate whether it is possible to improve sentiment analysis techniques on user-based reviews[5].

**1.3 Project scope**

This section will describe the scope of the project, and briefly outline the preconditions and constraints that narrows down the focus area of this thesis.

**1.3.1 Dataset**

We experimented with two different datasets in this research. We could use any set of movie reviews for the project, but we have chosen and limited it to movie reviews. In other words, we can say that we use movie reviews as an “use case” for our project .Our experiment, uses a corpus based on positive and negative movie reviews from IMDb. We gathered 1000 positive and 1000 negative reviews from the dataset “polarity dataset v2.0”6 first introduced in [5] in 2004 and another large dataset gathered 12500 positive and 12500 negative reviews from the dataset. This dataset is often used as a benchmark.

**1.3.2 Classification algorithms:**

The classification algorithms used in our experiments are limited to the Support vector machines and Multinomial Naïve Base algorithms. Support vector machines, are chosen because it is a classification algorithm common to use in the area of sentiment analysis. Multinomial Naïve Base is use to find best accuracy also support multi category.

* 1. **Tools/Technology :**

**1.4.1 Tools:**

* + - * Anaconda
      * Jupyter
      * Click chart diagram and flow chart software

**1.4.2 Language:**

Python 3.6

**1.4.3 Operating System:**

Windows 10

**1.5 Approach:**

This section will briefly describe how we proceeded to achieve our goal. We started with a restudy to examine the recommender and sentiment analysis approaches that exist today, what their functionality, possibilities and limitations[6]. The number of positive and negative were counted for each movie review and used as attributes for the classifier. Finally, the classification algorithms Support vector machines and Multinomial Naïve Base algorithms were used for classification. See Chapter 5 for a more detailed description of the approach.

**1.6 Main contributions:**

The main contributions of this project include an introduction to the basic principles behind todays recommender systems. We present the different approaches and challenges concerning the area of sentiment analysis. We present machine learning algorithms, including a detailed explanation of the Support vector machines and Multinomial Naïve Bayes. Most important, this thesis presents a system and an approach to how reviews can be used in the area of sentiment analysis. Based on the results proven in this thesis, our system can be used as a basis for a more powerful and personalized recommendation.

**1.7 Structure of report:**

The remainder of the report is structured as follows: First, Chapter 2 gives an introduction to the domain knowledge and the background of this thesis. Chapter 3 presents related work and how this thesis differs from work done by others. In Chapter 4 you get a detailed description of the classification methods used in this research. Chapter 5 describes the dataset used and the preprocessing of the same dataset. It also documents the approach of the research and how the results were collected. Chapter 6 presents the results from the experiments along with the evaluation metrics used. Lastly, Chapter 7 show the conclusions and presents improvements for future work.

**Chapter No 2**

**Literature Review**

**2. Literature Review**

**2.1. Sentiment Analysis and Opinion Mining**

Sentiment analysis and opinion mining is the field of study that analyzes people’s opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining. In fact, this research has spread outside of computer science to the management sciences and social sciences due to its importance to business and society as a whole. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, movie and social networks. For the first time in human history, we now have a huge volume of opinionated data recorded in digital form for analysis. Sentiment analysis systems are being applied in almost every business and social domain because opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are largely conditioned on how others see and evaluate the world. For this reason, when we need to make a decision we often seek out the opinions of others. This is true not only for individuals but also for organizations.[2] This book is a comprehensive introductory and survey text. It covers all important topics and the latest developments in the field with over 400 references. It is suitable for students, researchers and practitioners who are interested in social media analysis in general and sentiment analysis in particular. Lecturers can readily use it in class for courses on natural language processing, social media analysis, text mining, and data mining.[1]

**2.2 Online Paper Review Analysis:**

Sentiment analysis or opinion mining is used to automate the detection of subjective information such as opinions, attitudes, emotions, and feelings. Hundreds of thousands care about scientific research and take a long time to select suitable papers for their research. Online reviews on papers are the essential source to help them. The reviews save reading time and save papers cost. This paper proposes a new technique to analyze online reviews. It is called sentiment analysis of online papers (SAOOP). SAOOP is a new technique used for enhancing bag-of-words model, improving the accuracy and performance. SAOOP is useful in increasing the understanding rate of review's sentences through higher language coverage cases. SAOOP introduces solutions for some sentiment analysis challenges and uses them to achieve higher accuracy. This paper also presents a measure of topic domain attributes, which provides a ranking of total judging on each text review for assessing and comparing results across different sentiment techniques for a given text review. Finally, showing the efficiency of the proposed approach by comparing the proposed technique with two sentiment analysis techniques. The comparison terms are based on measuring accuracy, performance and understanding rate of sentences [7].

**2.3 Sentiment Analysis of Review Datasets using Naïve Bayes’:**

The advent of Web 2.0 has led to an increase in the amount of sentimental content available in the Web. Such content is often found in social media web sites in the form of movie or product reviews, user comments, testimonials, messages in discussion forums etc. Timely discovery of the sentimental or opinionated web content has a number of advantages, the most important of all being monetization. Understanding of the sentiments of human masses towards different entities and products enables better services for contextual advertisements, recommendation systems and analysis of market trends. The focus of our project is sentiment focused web crawling frame work to facilitate the quick discovery of sentimental contents of movie reviews and hotel reviews and analysis of the same. We use statistical methods to capture elements of subjective style and the sentence polarity. The paper elaborately discusses two supervised machine learning algorithms: K-Nearest Neighbor (K-NN) and Naïve Bayes and compares their overall accuracy, precisions as well as recall values. It was seen that in case of movie reviews Naïve Bayes’ gave far better results than K-NN but for hotel reviews these algorithms gave lesser, almost same accuracies [8].

**Chapter No 3**

**Background and Related Work**

**3. Background and Related Work**

We’ll discuss in chapter 3 about previous work done in that field and basic concepts used analysis (also known as opinion mining) refers to the use of natural language processing, text in this work. This chapter provides the reader with a fundamental understanding of the domain knowledge in the field of study.

**3.1** **Background Sentiment analysis**:

Sentiment analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader).

**3.2 Basic Concepts:**

**3.2.1.** **Sentiment feeling:**

Feelings, emotion, affection, sentiment, passion mean a subjective response to a person, thing, or situation. feeling denotes any partly mental, partly physical response marked by pleasure, pain, attraction, or repulsion, it may suggest the mere existence of a response but imply nothing about the nature or intensity of it (the feelings that once moved me are gone). emotion carries a strong implication of excitement or agitation but, like feeling, encompasses both positive and negative responses (the drama portrays the emotions of adolescence). Affection applies to feelings that are also inclinations or likings (a memoir of childhood filled with affection for her family). sentiment often implies an emotion inspired by an idea (her feminist sentiments are well known). passion suggests a very powerful or controlling emotion [9].

**3.2.2. Automated sentiment analysis:**

Automated sentiment analysis is the process of training a computer to identify sentiment within content through Natural Language Processing (NLP). Various sentiment measurement platforms employ different techniques and statistical methodologies to evaluate sentiment across the web. Some rely 100% on automated sentiment, some employ humans to analyze sentiment, and some use a hybrid system. Automated sentiment analysis will never be as accurate as human analysis, because it doesn't account for the subtleties of sarcasm or body language. if you have 1 to 10 movie review or may be 100, the most effective way to measure sentiment is to simply read them. But what happens if you have 50K? Or 100K. This is where automated sentiment can provide some directional insight and set the tone for further analysis. Like daily movie review about politics about consumer products is huge [5].

**3.2.3 Emoticons:**

An emotion icon, better known by the portmanteau emoticon is a meta communicative pictorial representation of a facial expression that, in the absence of body language and prosody, serves to draw a receiver's attention to the tenor or temper of a sender's nominal verbal communication, changing and improving its interpretation. It expresses — usually by means of punctuation marks (though it can include numbers and letters) — a person's feelings or mood, though as emoticons have become more popular, some devices have provided stylized pictures that do not use punctuation. As social media has become widespread, emoticons have played a significant role in communication through technology. They offer another range of tone and feeling through texting that portrays specific emotions through facial gestures while in the midst of text-based cyber communication [9].

**3.2.4 Features:**

The table tells us about the good and bad comment words are mostly in use for better judging that the sentence is positive or negative.

|  |  |
| --- | --- |
| **Positive** | **Negative** |
| good | bad |
| best | worse |
| Excellent | boring |
| Fascinating | violence |
| Dazzling | silly |

**Table 1 :** showing positive and negative words

There are many emotions words representing different expressions and meanings. Different people may have different preferences for using words. For instance, per the research results, people from Asian countries use the different emoticons to express their emotions in chat text messages compared with European people. Several new challenges due to the typical short length and irregular structure of such content. Two main research directions can be identified in the literature of sentiment analysis on microblogs. First direction is concerned with finding new methods to run such analysis, such as performing sentiment label propagation on movie review follower graphs, and employing social relations for user-level sentiment analysis. The second direction is focused on identifying new sets of features to add to the trained model for sentiment identification, such as microblogging features including hashtags, emoticons, the presence of intensifiers such as all-caps and character repetitions etc., and sentiment-topic features. Movie review has played an outsized role in a 2016 presidential election that continues to test the electorate. The ability of a single tweet to shape political conversation and drive media coverage has never been greater. The leading candidates for America's next presidency use Movie review to energize their supporters and draw citizens who wouldn't otherwise follow political discourse. Movie review simple and personal messages resonate in a way that more traditional means of communication mail rob calls and yard signs no longer can. Now Movie review plays major role in user’s opinion making. So, that’s why sentiment analysis is major tool for organizations to know about people views about their products, brands or policies to next time act per it or market per it for future products [10].

**3.4 Related Work Sentiment analysis:**

Sentiment analysis is a growing area of Natural Language Processing with research ranging from document level classification to learning the polarity of words and phrases (e.g... Given the character limitations on movie review, classifying the sentiment of Movie review messages is most like sentence-level sentiment analysis however, the informal and specialized language used in Movie review, as well as the very nature of the microblogging domain make Movie review sentiment analysis a very different Task. It’s an open question how well the features and techniques used on more well- formed data will transfer to the microblogging domain [8]. Just in the past year there have been several papers looking at Movie review sentiment. Other researchers have begun to explore the use of part-of-speech features but results remain mixed. Features common to microblogging (e.g., emoticons) are also common, but there has been little investigation into the usefulness of existing sentiment resources.

Developed on non-microblogging data. Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers rely on emoticons for defining their training data. Exploit existing Movie review sentiment sites for collecting training data, also use hashtags for creating training data, but they limit their experiments to sentiment/non-sentiment classification, rather than 3-way polarity classification, as we do [9].

**Chapter#4**

**System Overview**

**4. System Overview**

**4.1 Proposed System:**

We presented a system which fetches movie reviews form movie review dataset and then all movie review will assign polarity one by one by sentence level sentiment analysis and will attach polarity per appraisal theory.

The proposed recognition algorithm will classify the text of a sentence according to the following sentiment categories: Positive, Negative . The proposed algorithm will estimate sentiment weights for each sentimental category in the form of a numerical vector. The vector is used to determine the dominant sentiment type (the sentiment type with the highest weight) and the overall emotional valence of a sentence (is the emotion positive, negative). The suggested algorithm will tell us about which sentence is positive and which sentence is negative according to their correspondence. Our objectives are:

1. Pre-process the dataset.
2. Implement feature extraction by using count-vectorizer and tfidf transform.
3. Do sentiment analysis on extracted data by using Multinomial NB and SVM.
4. Give accuracy of algorithms by using Numpy.
5. Provide results by classification report.
6. Show results in graphical representation.

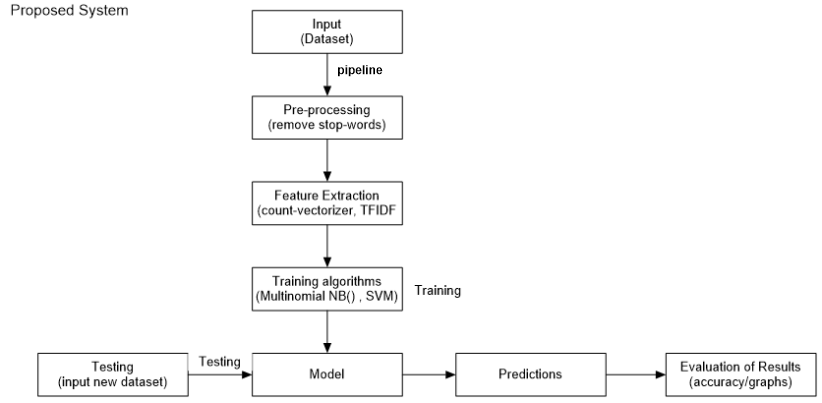


Fig 4.1 : Proposed System.

**4.2. Sentiment Recognition:**

**Sentiment Categories**:

Text should be categories in the following categories:

1. Positive.
2. Negative.

**4.3 Data cleaning:**

Sentiment analysis of movie review encounters several issues such as poor spelling, use of abbreviations, poor punctuation, poor grammar, use of slangs. Therefore, these hurdles need to be removed to ensure better results for sentiment analysis.

▪ Removal of stop words

**4.4 Non-Functional Requirements :**

**Usability:**

▪ the system’s interface is user friendly and easy to use. ▪ Any user with basic knowledge of PC can use it.

**Reliability:**

▪ Availability 99.9% The system shall accurately calculate the sentiments of user above 60%.

**Portability:**

▪ The system is run and tested on window 8.1 and 10.

**Performance:**

▪ The system will identify the sentiment approximately within 5 seconds.

**Chapter No 5**

**Experimental Setup**

**5.Experimental Setup**

This chapter describes our system in detail. The first section presents an overview of the system flow. The second section describes the details of the data collection and the third section explains the preprocessing of the datasets. The following section provides a detailed explanation of the approach and the system components. In the two last sections we explain how we prepared the data for classification along with the classification itself. The implementation of the system is mainly performed in Python.

**5.1 System flow:**

In this section we presents a high-level description of how our system is constructed: A collection of movie reviews, are preprocessed, involving stop-words. Further, we use feature extraction, to get the data in matrix form. Through this data can be classify easily in the classification process. Before we can start the classification, we need to transform the results into a format that is ready for classification. The following sections will describe each of these components in greater detail. The system flow of this project is divided into the following steps:

1. **Obtaining the dataset:**
2. Using scikit-learn.

**2.** **Preprocessing of the data:**

(a) stop-words removal.

**3. Feature Extraction:**

(a) Count Vectorizer.

(b) TF-IDF Transformer.

**4. Algorithms:**

1. Multinomial NB().
2. SVM.

**5. Prediction:**

1. Numpy.

**5.1.1 Obtaining the dataset:**

The dataset which is taken from corpus of movie reviews has 2000 review which are categorized in positive and negative. 1000 reviews are positive and 1000 reviews are negative which are used for training the algorithms as show below:

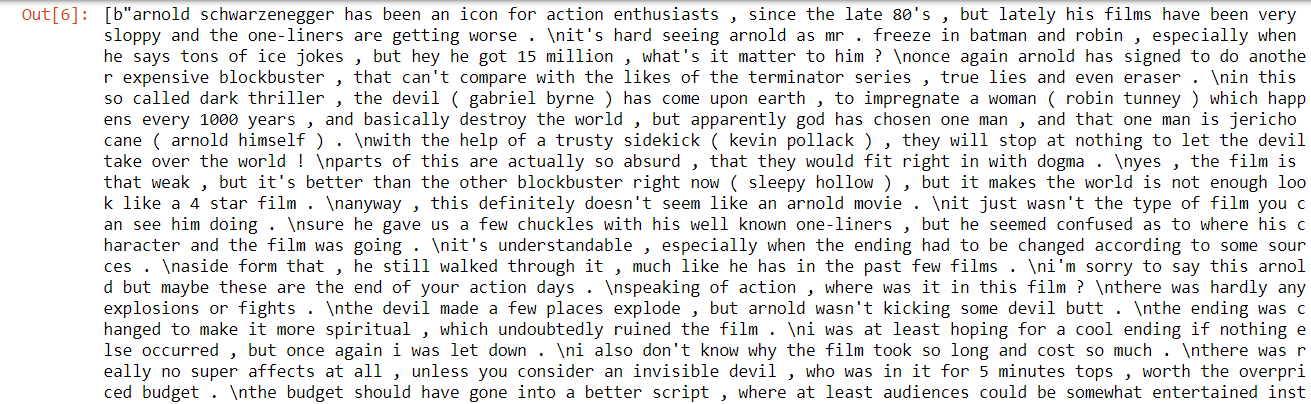


Fig 5.1 : Raw data.

**5.1.1.2 Scikit Learn:**

Scikit learn is a free software machine learning library for the python programming language. We use scikit learn library to load the dataset. As you can see fig 5.1 a



Fig 5.1a : Loading dataset by using library.

**5.1.2 Preprocessing of the data:**

Before information can be extracted from the data, it must be preprocessed to prepare it for further processing. There are many operations that can be applied for this task, including elimination of stop words, stemming, lemmatization and part-of-speech tagging. This section will explain the techniques we find relevant for this research.

**5.1.2.1 Stop-words removal:**

Stop words are the words that are use to purify the data during pre-processing. Stop words are the most common words in the language. There is no single word in the universe which is use by all natural language process tools. Some tools specify to avoid the stop words. There are some common words such that “ the , is , at , that , etc ’’. [11]



Fig 5.2 : Stop-words.

These are the stop-words which are necessary removed from the dataset to clean the data to get better accuracy. But if, we use stop-words library to clean the data then it will give us the better results and give us clean data.

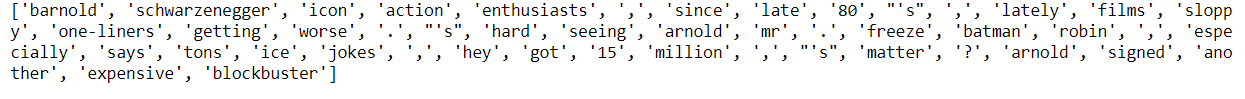


Fig 5.2.1 : After removing stop-words.

**5.1.3. Feature selection:**

**5.3.1 Count-Vectorizer:**

A collection of text documents is converted into a token count of matrix. When we apply Count-Vectorizer then it will produce the sparse matrix of the counts and change the text data into matrix form. [12]

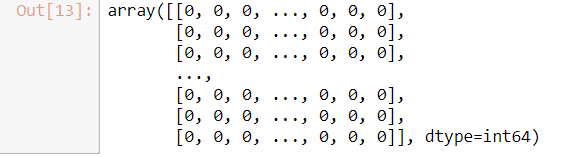


Fig 5.3 : clean data changes into matrix form.

**5.3.2 TF-IDF Transform:**

A standard approach to represent documents is by use of vectors, where each dimension represents one feature. The terms are often weighted and in this research, we used the TF-IDF weighting method. TF stands for term frequency and is the number of times the term occurs in the specific document. In other words, TF tells us how widespread the term is in the document and the more often a term occurs in in a document, the better the term describes the document. DF is the number of documents in the whole collection that contains the term. A term that occurs in almost every document is not helpful in discriminating between classes. Typical words can be “the”, “and”, “as” and so on. We want terms that occur in fewer documents to be weighted higher and this is what inverted document frequency IDF does. The inverted document frequency is defined as follows:

IDF = log |N| / DF

where N is the total number of documents in the collection and DF is the document

frequency. By multiplying TF with the IDF we get a weight that increases with many

occurrence of a term within a document and decreases with the number of documents

that contain the term. The TF-IDF is defined as TF multiplied with IDF as follows: [12]

TF-IDF = TF \* IDF Where , IDF = log |N| / DF.

TF-IDF = TF \* log |N| / DF

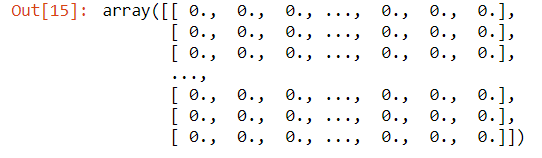


Fig 5.4: Sparse matrix data is much clean by TF-IDF.

**5.1.4 Algorithms :**

An algorithm in data mining ( or machine learning) is a set of process and estimation that creates a model from data. To create a model :

1. The data which you provide will firstly analyzes by the algorithm.
2. Look for the distinct type of patterns and trends.

The algorithm gets the results and uses many iteration to find the optimal solution for creating the mining model [13]. Here, two algorithms will be used one is Multinomial Naïve Bayes and another is SVM (Support Vector Machine).

**5.4.1 Multinomial NB() :**

Multinomial Naïve Bayes is the supervised learning algorithm. It implements the naïve bayes algorithm for distributed the data by multinomial. It is one of the two classic naïve bayes variants used in the text classification ( where the data is typically represent as word vector counts and tfidf vectors are also work as well). The mathematical formula is :

P(c) = Nc / N

P(w/c) = count(w,c) +1 / count(c) + |V|

Where,

* P(c) is the probability of class.
* N(c) is the number of document in the class.
* N is the total number of document.
* w is words.
* c is class.
* V is vocabulary size.
* Count (w,c) is that how many times the word comes.
* Count(c) is that how many total words are there in the training set.

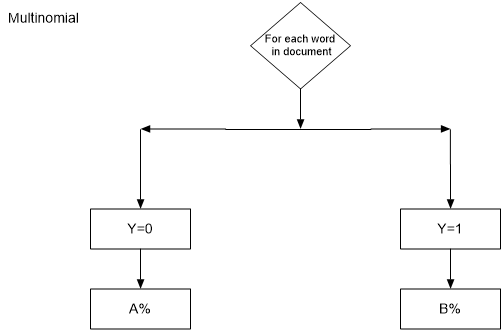


Fig 5.4: Multinomial NB() classifier

**Where to use:**

The naïve bayes implementation can be found in Orange , Sci-ket learn , Weka and R.

**5.4.2 Support Vector Machine:**

SVM is a vector-based classification method. It is known as a fast and very effective algorithm for classifying the problems. We can think of SVM as a hyperplane in the feature space. The hyperplane is chosen during training the data. The hyperplane is that line which separates the positive from negative with maximal margin. The margin is defined as the distance from hyperplane to the nearest point from positive and negative sets. SVM hyperplanes are strong-minded by a relatively small subset of the training occurrences called, as the support vectors. The respite of the training data have no impact on the trained classifier. Figure 5.4 will shows a highest margin hyperplane in two dimensions [14].

**Where to use:**

There are many implementations of SVM. Most popular ones are scikit-learn , MATLAB and libsvm.

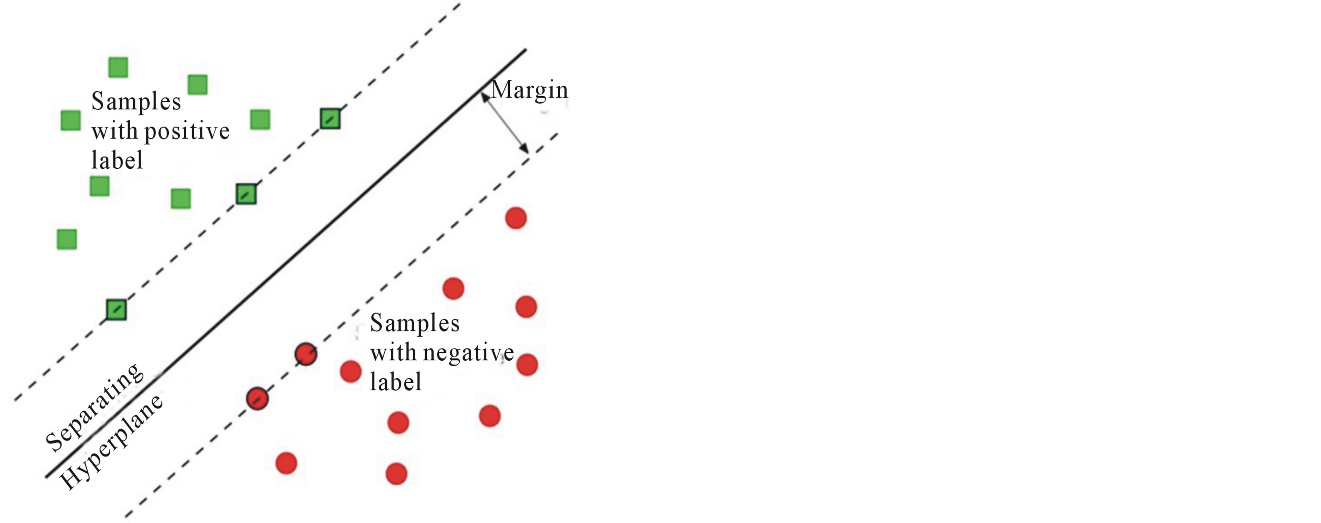


Fig 5.4 : SVM classifier in two dimensions.

**5.5 Prediction:**

Prediction in data mining is to detect the data points only on the report of another data rate value. It is not certainly related to future events but the used variables are unknown. Predictions efforts the relationship between a thing you know, and a thing you need to predict for future orientations [17].

**5.5.1 Numpy:**

Numpy is the basic library for scientific computing in python. It offers a high- performance multi-dimensional array object and tools for working with these arrays. It is use in this project to get the accuracy of algorithms of the dataset.

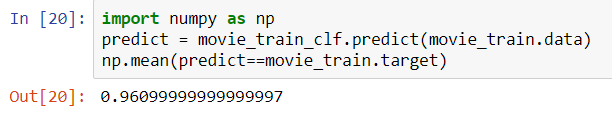
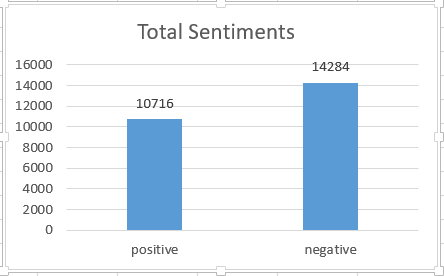


Fig 5.5: Using numpy gives accuracy.

It is also use for counting the number of positive and negative reviews. Like, in our dataset , you will see number of positive reviews and negative reviews.



**Chapter 6**

**Results and Discussion**

**6. Results and Discussion**

Before presenting the results, this chapter describes about the evaluation matrix of the experiments . The results includes both the confusion matrices and detail accuracy of the classification also discussion about them.

**6.1 Evaluation Matrices:**

This chapter will explain the confusion matrices along with the evaluation matrices: “TP Rate/Recall” , “FP rate” , “Precision” , “F-measure” and “Accuracy”. These concepts are important. So, you will be able to understand the results.

A **Confusion matrix** is also called Error matrix. A matrix (table) that can be used to measure the performance of an machine learning algorithms, usually a supervised learning one. Each row of the confusion matrix represents an actual class and each column represents the predicted class. The table 6.1 shows an example of an confusion matrix. The correct classification is marked with bold text.[15]

|  |  |  |
| --- | --- | --- |
| A | B | <= classified as |
| **True class: a**  **Predicted class: a** | True class: a  Predicted : b | a = positive |
| True class: b  Predicted class: a | **True class: b**  **Predicted class: b** | b = negative |

**Table 6.1**: Example of Confusion Matrix

TP stands for **True Positive.** A true positive means a correctly classified positive document. For example , if a document is classified positive and the “true sentiment” also is positive, the document is countered as a True Positive. The **TP rate** also called **Recall**, measures the proportion of positives that are correctly identified by the classifier and is calculated as follows:

TPRate = Recall = TP / P = TP / TP + FP. (6.1)

Here FN means a False Negative. A false negative is a document that is returned as negative by the classification, but in reality is positive. Opposite of the FN is the False Positive FP , FP occurs if a document is negative and the classification returns a positive result. In other words, the FP Rate defines how many incorrect positive results that occur among all the negative samples and can be calculated as:

FPRate = FN / N = FN / FN + TN. (6.2)

TN represents a true negative; the classification correctly classifies a negative document. A summary of what TP , FP , FN and TN means, is presented in Table 6.2.

|  |  |  |
| --- | --- | --- |
|  | True sentiment positive | True sentiment negative |
| Classified Positive | TP | FP |
| Classified Negative | FN | TN |

**Table 6.2 :** Description of “True Positive” , “False Positive” , “False Negative” , “True Negative”

An often used method to calculate classification tasks is **precision**. Precision tells us which fraction of the documents that are correctly classified:

Precision = TP / TP + FN (6.3)

**F-measure** is a measure of a test’s accuracy and is the harmonic mean of precision and recall. It has a parameter that sets the tradeoff between recall and precision. The standard F-measure F1 assigns equal importance to recall and Precision, and is defined as follows:

F1 = 2 \* Precision \* Recall / Precision + Recall = 2TP / 2TP+FN+FP (6.4)

The **F1-score** can be showed as the weighted average of the precision and recall. The F1 score therefore reaches its best value at 1 and it’s worst value at 0. The general formula of a real positive number β is defined as:

Fβ = (1 + β2) Precision∗Recall **/** (β2 ∗Precision) + Recall (6.5)

**Accuracy** is the proportion of the total number of predictions that were correct. Be careful , if the dataset is very unbalanced , accuracy is not a reliable metric for the real performance of a classifier. For example, if we have 95% positive reviews and 5% negative reviews and the classiﬁer predicts all the instances to be positive, we will get an overall accuracy of 95%. In practice the classiﬁer have a 100% recognition rate for the positive reviews and 0% recognition rate for the negative reviews. If this problem should occur, it would be easily detected by the calculation of the other metrics. The accuracy is determined using the equation: [16].

Accuracy = TP + TN / Total = TP + TN / TP + FP + FN + TN (6.6)

**6.2 Results by Multinomial NB() on the small dataset:**

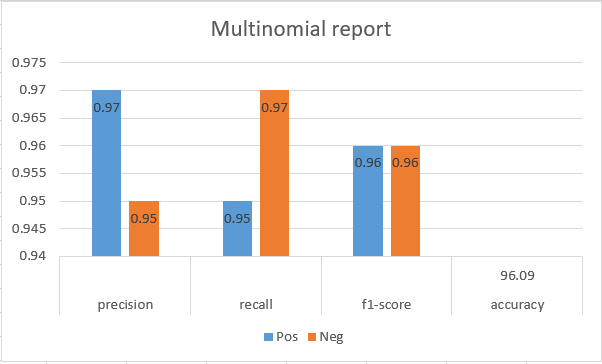
**Table 6.3** : Confusion matrices of Multinomial NB()

|  |  |  |
| --- | --- | --- |
| a | B | <= classified as |
| 974 | 26 | a = positive |
| 52 | 948 | b = negative |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Precision** | **Re-call** | **F1-score** | **Support** | **Accuracy** | **Class** |
|  | 0.97 | 0.95 | 0.96 | 1000 | 0.9609 | pos |
|  | 0.95 | 0.97 | 0.96 | 1000 | 0.9609 | neg |
| **Weighted avg.** | 0.96 | 0.96 | 0.96 | 2000 | 0.9609 |  |

**Table 6.4** : Detail accuracy by class of Multinomial NB()

It can also be shown graphically:



**6.3 Results by Support Vector Machines on the small dataset:**

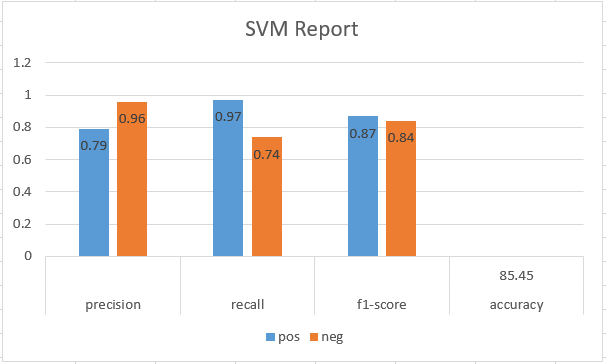
**Table 6.5 :** Confusion matrices of SVM

|  |  |  |
| --- | --- | --- |
| a | B | **<=** classified as |
| 739 | 261 | a = positive |
| 29 | 971 | b = negative |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** | **Accuracy** | **Class** |
|  | 0.79 | 0.97 | 0.87 | 1000 | 0.8545 | pos |
|  | 0.96 | 0.74 | 0.84 | 1000 | 0.8545 | neg |
| **Weighted avg** | 0.88 | 0.85 | 0.85 | 2000 | 0.8545 |  |

**Table 6.6** : Detail accuracy by class of SVM.

It can be presented graphically:



**6.3 Results by Multinomial NB() on the large dataset:**

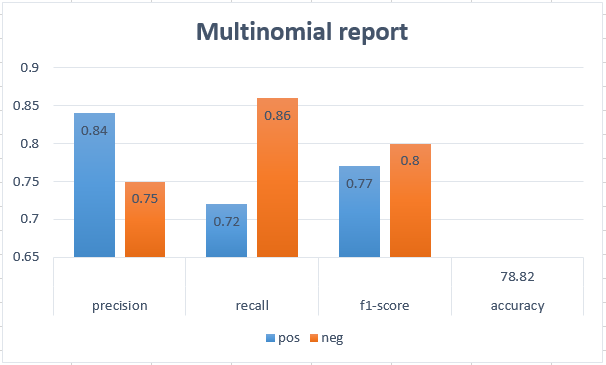
**Table 6.7 :** Confusion matrices of Multinomial NB()

|  |  |  |
| --- | --- | --- |
| A | b | <= classified as |
| 10745 | 1755 | a = positive |
| 3539 | 8961 | b = negative |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** | **Accuracy** | **Class** |
|  | 0.84 | 0.72 | 0.77 | 12500 | 0.7882 | pos |
|  | 0.75 | 0.86 | 0.80 | 12500 | 0.7882 | neg |
| **Weighted avg.** | 0.79 | 0.79 | 0.79 | 25000 | 0.7882 |  |

**Table 6.8** : Detail accuracy by class of Multinomial NB().

It can be represent graphically:



**6.4 Results by SVM on the large dataset:**

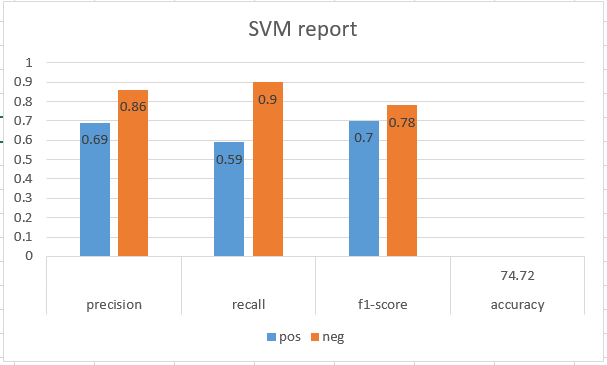
**Table 6.7 :** Confusion matrices of SVM

|  |  |  |
| --- | --- | --- |
| A | b | <= classified as |
| 7427 | 5073 | a = positive |
| 1245 | 11255 | b = negative |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** | **Accuracy** | **Class** |
|  | 0.69 | 0.90 | 0.78 | 12500 | 0.7472 | pos |
|  | 0.86 | 0.59 | 0.70 | 12500 | 0.7472 | neg |
| **Weighted avg.** | 0.77 | 0.75 | 0.74 | 25000 | 0.7472 |  |

**Table 6.8** : Detail accuracy by class of SVM.

It can be shown graphically:



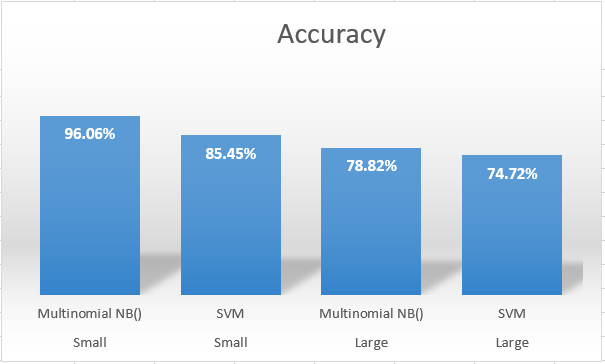
**6.5 Accuracy of small and large dataset :**

In this chapter, we will also do discussion about the accuracy of our algorithms. Here, we use two algorithms Multinomial NB() and Support Vector Machine (SVM). Now, we will going to show the accuracies of our results on small dataset and large dataset.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Algorithms** | **Accuracy** |
| Small | Multinomial NB() | 96.06% |
| Small | SVM | 85.45% |
| Large | Multinomial NB() | 78.82% |
| Large | SVM | 74.72% |

**Table 6.9:** Accuracy in all our Experiments

In table 6.9 , by using small dataset , Multinomial NB() gets 96.06% accuracy instead of SVM. Similarly, by using large dataset , Multinomial NB() get 78.82% accuracy instead of SVM. So, the results in small and large dataset is that Multinomial NB() is much better than the SVM. You can see it’s graph also which will give much better clarity.



**Chapter 7**

**Conclusions**

**7. Conclusions:**

This chapter concludes the work presented in thesis , including the evaluation. The purpose of this project was clearly done by telling whether the review is positive or negative. We also use pre-processing which helps us out to clean much of our data. Then we use two algorithms and train on small dataset and test on large dataset. After that we see that Multinomial Naïve Bayes algorithm is much better than Support Vector Machines algorithm by getting accuracy 96.09% on small dataset. When we do testing then we observe that Multinomial Naïve Bayes algorithm is better than Support Vector Machine algorithm with accuracy 78.82% on large dataset. So we conclude that Multinomial Naïve Bayes algorithm is very effective and good in training and testing of datasets.

We believe that research within the sentiment analysis and recommender system will become more important each day as the user-based reviews increases in numbers on the Internet. It forms solid basis for future research.

**Future work:**

We experienced trouble of finding out sentiments on reviews , because some sentence has both positive and negative reviews which could be trouble for system to judge. For future work , I pursue to get much more further and experience the text dataset on multiple classes. As for sure there comes much impurity in the dataset which we have to clean out and try it on multiple classes which will be a great progress in this field and people learn a lot about it. Making this system will surely help user for recommending a great movie to watch with strong predictions.

**References:**

**[1]** Zhuang, Li, Feng Jing, and Xiaoyan Zhu. *Movie review mining and summarization*. In Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006).

**[2]** 1Zirn, Cäcilia, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. *Fine-Grained* Sentiment Analysis with Structural Features. In Proceedings of the 5th International Joint Conference on

Natural Language Processing (IJCNLP-2011). 2011.

**[3]** A. Dragland, “Big data – for better or worse,” SINTEF, vol. 22, May 2013.

**[4]** P. Melville and V. Sindhwani, “Recommender systems,” in Encyclopedia of Machine

Learning, pp. 829–838, 2010.

**[5]** J. A. Horrigan, “Online shopping. pew internet & american life project report,”

2008.

**[6]** C. A. Gomez-Uribe and N. Hunt, “The netflix recommender system: Algorithms,

business value, and innovation,” ACM Trans. Manage. Inf. Syst., vol. 6, pp. 13:1–

13:19, Dec. 2015.

**[7]** B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity

summarization based on minimum cuts,” in Proceedings of the ACL, 2004.

**[8]** D. Bamman, B. O’Connor, and N. A. Smith, “Learning latent personas of film

characters.,” in ACL (1), pp. 352–361, The Association for Computer Linguistics,

2013.

**[9]** A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, “Learning

word vectors for sentiment analysis,” in Proceedings of the 49th Annual Meeting

of the Association for Computational Linguistics: Human Language Technologies,

(Portland, Oregon, USA), pp. 142–150, Association for Computational

Linguistics, June 2011.

**[10]** K. Balog, N. Takhirov, H. Ramampiaro, and K. Norvaag, “Multi-step classification

approaches to cumulative citation recommendation,” in Open research Areas

in Information Retrieval (OAIR 2013), 2013.

**[11]** https://en.wikipedia.org/wiki/Stop\_words

**[12]**http://scikitlearn.org/stable/modules/generated/sklearn.feature\_extraction.text.CountVectorizer.html

**[13]** https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-analysis-services-2017

**[14]** <https://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html/3>

**[15]** <https://www.python-course.eu/confusion_matrix.php>

**[16]**<http://scikitlearn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html>

**[17]** https://www.quora.com/What-is-prediction-in-data-mining