PREMIER UNIVERSITY, CHATTOGRAM
Department of Computer Science & Engineering

AIL Project Report

On

Sentiment Analysis of IMDB Movie Reviews

SUBMITTED BY

Name: Mohiuddin Tamim

ID: 1903610201763

Name: Ami Biswas

ID: 1703310201400

Name: Zubayer Bin Rashid

ID: 1903610201782

Course instructed

by

Faisal Ahmed

Lecturer

Department of Computer Science & Engineering

Premier University, Chattogram

September, 2022

Author's Declaration of Originality

We hereby declare that the report work entitled **"Sentiment Analysis of IMDB Movie Reviews"** submitted to the Premier University, is a record of an original work done by us under the guidance of Mr. Faisal Ahmed, Lecturer, Department of Computer Science & Engineering, Premier University, Chittagong. We can assure that the result of this report has not been submitted to any other university.

Mohiuddin Tamim
ID: 1903610201763

Ami Biswas
ID: 1703310201400

Zubayer Bin Rashid
ID: 1903610201782

CERTIFICATION

The report entitled **"Sentiment Analysis of IMDB Movie Reviews"** submitted by, Mohiuddin Tamim, ID: 1903610201763, Ami Biswas, ID: 1703310201400, Zubayer Bin Rashid, ID: 1903610201782 has been accepted as satisfactory in fulfillment of the course Artificial Intelligence Lab.

Faisal Ahmed
Lecturer
Department of Computer Science & Engineering
Premier University, Chattogram

Table of Contents

List of Figures

List of Tables

Abstract

Sentiment analysis has become essential for creating opinion mining systems because of the massive volume of data and opinions that are created, shared, and moved every day through the internet and other media. It focuses on analyzing and understanding the emotions implied by textual patterns. It characterizes the expressing of feelings, such as those that are neutral, positive, or negative towards the existence of something, instinctively. Data analysis can employ a variety of sources, including reviews of movies, social media, newspapers, and medical information. Here, we have collected IMDB movie review data as well as used five kinds of machine learning classifiers to analyze these data. Hence, the considered classifiers are Gaussian Naïve Bayes (BNB), Decision Tree (DE), Logistic Regression (LR) as well as K-Nearest Neighbours (KNN). According to our analysis, DT displays higher precision when compared to others, while LR achieves better accuracy, precision, and F-score.

Keywords: Sentiment Analysis, Machine Learning, Natural Language Processing, Tokenization, Stemming, IMDB Reviews, Movie Reviews, Classification, Naive Bayes, k-Nearest Neighbors, Decision tree, Logistic Regression.

CHAPTER 1

INTRODUCTION

1.1 Background

The easiest way for individuals to find enjoyment is through movies. A small number of films, however, are well-received and popular. There are numerous rating services that can assist movie enthusiasts in deciding which films to view and which to skip. Among those, the most popular websites include IMDB, Rotten Tomatoes, etc. According to the stars that spectators have given the film, these websites' ratings measure its viability by assigning it a score out of 10. However, no technology exists that can make a prediction based on movie reviews. Sentiment analysis is thus used to assess the movie's performance based on reviews.

Sentiment analysis, commonly referred to as opinion mining, examines how people feel about various objects, including goods, services, groups of people, organizations, topics, events, and their characteristics. The decision-making process of recommendation systems has historically given significant weight to "the thoughts, opinions, and choices of others," as seen by the fact that recommendations for products on Amazon and Walmart are based on sentiment analysis of user reviews.

Sentiment analysis is the process of determining how the general public feels about a particular film and how they will evaluate it. For more information, the user evaluations' documents are categorized according to the attitude they are conveying, including favorable, somewhat positive, somewhat negative, and negative. Text mining, Natural Language Processing (NLP), and other computational methods are used in sentiment analysis.

Tokenization, word filtering, stemming, and classifications are all part of the sentiment

analysis methodology. Tokenization requires the division of text into discrete elements like words, integers, or punctuation. The next stage is stemming, which is the act of eliminating prefixes and affixes to reveal a word's stem. After preprocessing, we perform classification on the dataset using Nave Bayes, KNN classification, Decision Trees, and Logistic Regression. Here, we choose the most accurate model. As a result, we evaluate and research the aspects that have an impact on the ratings of our review text before classifying the movie as good or negative.

1.2 Motivation

Sentiment analysis is the evaluation and categorization of emotions found in text data using text analysis methods. Businesses can determine customer sentiment through online discussions and comments by using sentiment analysis. Sentiment analysis models concentrate on polarity (positive, negative, neutral), but also on sentiments and emotions (angry, pleased, sad, etc.), and even on intentions (e.g. interested v. not interested). Sentiment analysis has gained popularity, and many major corporations are devoting resources to it in an effort to forecast outcomes for their operations.

1.3 Objective

In the operation of sentiment analysis tokenization, word filtering, stemming, text representation, and classifications all play a vital role. Text must be divided into tokenized units like words, numbers, or punctuation. The next stage is stemming, which involves stripping a word of its prefixes and affixes to reveal its stem. Then, text documents are converted to matrix using TFIDF, a text representation. Next, we classify the dataset using Naive Bayes, K-Nearest Neighbor Classifier (KNN), Decision Tree, and Logistic Regression. In this case, accuracy is used to select the best model. In order to classify the movie as favorable or negative, we examine and research the characteristics that have an impact on the scores of our review text.

1.4 Summary

In this report, we present a model to find the sentiment analysis of IMDB movie reviews. When applying text analysis tools, sentiment analysis interprets and categorizes the emotions included in text data. Here many stages will be followed, which are tokenization, stemming and after preprocessing we will perform classification on the dataset. Afterwards, we will be able to choose the most accurate model

CHAPTER 2

LITERATURE REVIEW

2.1 Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory

The contributor of this journal is Saeed Mian Qaisar
College of Engineering, Effat University, 21478, Jeddah, Saudi Arabia

For this paper they used a data set containing 50k movie reviews from IMDB, created by Andrew Maas.
Each movie review is encoded "vectorized" into a numeric value using Natural Language Processing and topic modeling utilizing genism library of python, Doc2Vector model, Long Short-Term memory (LSTM) Classifier. The highest accuracy attained by the devised approach is of 89.9 percent.
A superior accuracy can be attained by using further data preconditioning techniques. Furthermore, higher classification accuracy can be achieved by employing the ensemble classifiers or deep learning approaches. [1]

2.2 Learning Word Vectors for Sentiment Analysis

The contributors of this journal are Andrew L. Maas, Raymond E. Daly, Peter T. Pham,
Dan Huang, Andrew Y. Ng, and Christopher Potts
Stanford University Stanford, CA 94305

For this paper, they used a large dataset of informal movie reviews from the Internet Movie Database (IMDB)
They used Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words, Document Polarity Classification. They got 50 percent accuracy here.
They extended the unsupervised model to incorporate sentiment information and demonstrated the utility of such representations on two tasks of sentiment classification, using existing datasets as well as a larger one that they will release for future research. These tasks involve relatively simple sentiment information, but the model is highly flexible in this regard; it can be used to characterize a wide variety of annotations, and thus is broadly applicable in the growing areas of sentiment analysis and retrieval. [2]

2.3 Movies Reviews Sentiment Analysis and Classification

The contributors of this journal are Mais Yasen, Sara Tedmori
Department of Computer Science
Princess Sumaya University for Technology
Amman, Jordan

In this paper, they used real reviews dataset from IMDB which contained almost 43 thousand instances for training and testing.
They used tokenization using NLTK, word filtering, stemming using Porter Stemming Algorithm, attribute selection using Gain Ratio Algorithm, Naive Bayes (NB), Decision Tree (DT), Support Vector Classifier (SVM), Bayes Network (BN), K-nearest Neighbors (KNN), Ripper Rule Learning (RRL), Random Forest (RF), Stochastic Gradient Descent (SGD). Here Random Forest got the best accuracy of 96.01 percent in comparison with all other classifiers.
The authors wish to conduct a similar study on different languages specifically on Arabic. In addition, the authors wish to experiment with different SA methods in order to increase the accuracy of the results. [3]

2.4 Comparative Study on Sentiment Analysis on IMDB Dataset

The contributors of this journal are Debarghya Banerjee, Sreya Mazumder, Samik Datta
Department of Computer Science and Engineering, Techno Engineering College
Banipur, West Bengal, India

For the sentimental analysis on IMDB movie reviews, they have collected the input data from Kaggle.
They used Bag of words, Term Frequency-Inverse Document Frequency (TF-IDF), K-Nearest Neighbour Classifier (KNN), Decision tree classifier, Random Forest classifier, Logistic Regression (LR), Support Vector Machine (SVM). The accuracy of Logistic Regression (LR) is 74.13 percent.
In future, a comparison of the performance analysis of deep learning algorithms will be conducted and that might result in the improvement of the accuracy of the models. [4]

2.5 Sentiment Analysis of Movie Reviews using Machine Learning Techniques

The contributors of this journal are Palak Baid, Apoorva Gupta, Neelam Chaplot
Department of Computer Science & Engineering Jaipur Engineering College and
Research Center Jaipur, Rajasthan,India

The data was collected from 2000 user-created movie reviews archived on the IMDb (Internet Movie Database) web portal at `http://reviews.imdb.com/Reviews` and is known as "Sentiment Polarity Dataset version 2.0": 1000 positive and 1000 negative processed reviews.
They used the StringToWordVector filter, Three algorithms are performed on the data generated, algorithms are Naïve Bayes, K Nearest Neighbour, and Random Forest. Naive Bayes performed best with 81.4 percent accuracy.
As only few algorithms were tested, it is required to test other algorithms or create hybrid methods so that accuracy of the results can be increased. [5]

2.6 Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification

The contributors are V.K. Singh, R. Piryani, A. Uddin, P. Waila
Department of Computer Science South Asian University
New Delhi, India

The data was collected from 10 reviews each for 100 Hindi movies from the popular movie review database website `www.imdb.com`. We have labeled all these reviews manually to evaluate performance of our algorithmic formulations. Out of 1000 movie reviews collected, 760 are labeled positive and 240 are labeled as negative reviews.
They used three algorithms to perform on the data generated SWN (AAC), SWN (AAAVC) and Alchemy API. SWN (AAAVC) performed best with 78.7 percent accuracy.
The only restriction with this aspect-level implementation is that it is domain specific. However, only a few changes (in aspect vectors) would be required to use this algorithmic formulation in a different domain. [6]

2.7 Sentiment Analysis On Movie Reviews Using NAÏVE BAYES Classifier

The contributors of this journal are Mohana Pranadeep potti, ManneDineshKmar,
Nagabhyrava Saswanth Ram,P.V.R.Sandeep, P.R.Krishna Prasad
Computer Science Department, Vasireddy Venkatadri Institute of technology,
Andhrapradesh, India

The dataset has been created from Twitter posts of movie reviews and related tweets about those movies.
They used NAÏVE BAYES. It performed with 85.1 percent accuracy.
They didn't use any other algorithm. Since their results were very close on Naive Bayes, they could have used other algorithms to see if those algorithms could get better results. [7]

2.8 Sentiment Analysis of IMDB Movie Reviews

The contributors of this journal are Abhimanyu Singh, Chaitanya Kulkarni, Necati A. Ayan
SUNY Binghamton

The gathered dataset from the publication of Abdalraouf Hassan and Ausif Mahmood that has 50000 reviews from IMDB which is equally divided into 25000 for training and testing.
They have used wordnet, porter stemmer lemmatization, tokenization using NLTK library, Bag of Words (BOW), Naïve Bayes, Logistic Regression, Support Vector Machine. In Logistic Regression, with Countvectorizer they got an accuracy score of 86.89 percent.
In future, they will conduct a comparison of the performance analysis of deep learning algorithms and that might result in the improvement of the accuracy of the models. [8]

2.9 Sentiment Analysis For Movies Reviews Dataset Using Deep Learning Models

The contributors of this journal are Nehal Mohamed Ali, Marwa Mostafa Abd El Hamid and Aliaa Youssif
Faculty of Computer Science, Arab Academy for Science Technology and Maritime
Cairo, Egypt

They used the IMDB dataset consisting of 50K movie review files(25K positive review files and 25K negative review files) reviews, all written in English. Files sizes have ranged from 1kb-15kb. No rating information was included in the text files. Dataset was split into 80 percent training set and 20 percent testing set.
They used deep learning models MLP, CNN, LSTM, CNN LSTM SVM, SNN LSTM, NB and RNTN. CNN LSTM performed best with 89.20 percent accuracy. [9]

2.10 Deep learning for sentiment analysis of movie reviews

The contributors of this journal are Hadi Pouransari, Saman Ghili
Stanford University

They use the dataset Kaggle which is publicly available: The labeled data set consists of 50,000 IMDB movie reviews, specially selected for sentiment analysis.
Three algorithms are performed on the data generated, algorithms are Random Forest, Logistic regression and SVM. Logistic regression performed best with 86.6 percent accuracy. The only restriction with this aspect-level implementation is that it is domain specific. However, only a few changes (in aspect vectors) would be required to use this algorithmic formulation in a different domain. [10]

2.11 Summary

In this section, we gathered information about 10 journals. We described about their dataset, the classifiers and everything they have used to train, run and optimize the model with the best accuracy score. Afterwards, we wrote about the limitation and future works of their work.

CHAPTER 3

METHODOLOGY/PROPOSED METHOD

3.1 Dataset Description

For this paper a data set containing 50k movie reviews from IMDB, created by Lakshmipathi N is utilized. Only 10,000 reviews were used in this research. Using the IMDB rating system, the reviews are divided into positive and negative categories. It allows users to rate on a scale from 1 to 10, with everything below 4 stars being tagged as bad and anything over 7 stars being recognized as positive, according to the dataset's developer. Reviews with ratings outside of the aforementioned ranges are not taken into account. Each movie has a maximum of 30 reviews. With a standard deviation of 172.91 words, reviews typically include 234.76 words. The collection has 46647 terms in total. Glimpse of the dataset positive and negative reviews is shown in Table 3.1.

No	Review	Sentiment
0	One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked. The...	positive
1	I saw this movie when I was about 12 when it came out. I recall the scariest scene was the big bird ...	negative
2	The story is about a psychic woman, Tory, who returns to her hometown and begins reliving her trauma...	negative

Table 3.1. Glimpse of the dataset positive and negative reviews.

In the dataset we acquired two types of sentiments positive or negative. The positive sentiments are 5028 in number and negative sentiments are 4972 in number. The graphical view of both the sentiments are shown in Fig.3.1

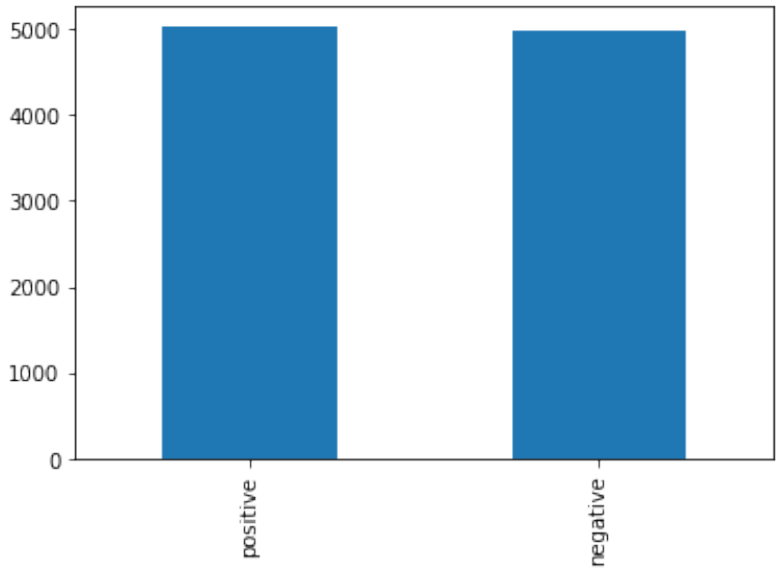

Figure 3.1. Graphical view of positive and negative sentiments.

3.2 Preprocessing

False identifications could result from poorly structured and unclean data. As a result, one of the most important steps in the data mining process is preparing the data.
Cleaning up the data from unused information that won't aid in training and could be confusing during the categorization process is referred to as this. Multiple data preparation techniques are applied to the IMDB dataset.

- **Removing special characters:** As we're working with English-language evaluations in our dataset, we need to make sure that any special characters (all exclamation points and question marks) are eliminated. [11]
- **Case Conversion:** The "lower case letters" are mostly viewed when compared with the uppercase letters due to the common utilization. Thus, the conversion of the uppercase letters to the lowercase letters is necessary and provide a simpler classification of the sentiments in this proposed system. [11]
- **Stop word removal:** It is employed for removing the most often used words like adverbs, conjunction, prepositions and article, which results in reduction of the dimensionality of the datasets. Some of the words are 'they', 'she', 'but', 'he', 'if' and 'we', so on. Thus, the stop words must be eradicated to enhance the quality of

sentiment analysis. [11]

- **Lemmatization:** It is a process of converting the given word to its root word. The main objective of lemmatization is to get proper morphological meaning of a word by referring it to the dictionary which is incorporated in the library. We have used wordnet and porter stemmer lemmatization [11]

3.3 Text Representation

A common method for weighing terms in NLP tasks is called TF-IDF (term frequency-inverse document frequency). The TF-IDF weights words based on their importance, therefore it can be used to determine which words are the most crucial. Therefore, some NLP tasks that make use of TF-potent IDF's weighting operation include text summarization, information retrieval, and sentiment categorization. This can be used to find the keywords (or even tags) for a document, or to more rapidly summarize articles.

3.4 Classification

We used K-Nearest Neighbor Classifier[3.4.2], Logistic Regression[3.4.4], Decision Tree[3.4.3], and Naive Bayes[3.4.1] in our experiment. In order to predict whether a movie would be favorable to watch or not, we have trained our model using the aforementioned classifiers.

3.4.1 Naive Bayes Classifier

The simplest and best-known classifier, NB, considers each attribute to be unique from the others. NB classifiers are a group of classification algorithms, not a single algorithm but a family of algorithms. The mathematical expression is as follows:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Where

$$P(Y|X) = P(y_1|X)P(y_2|X)...P(y_n|X)$$

Here, X is class variable and Y is a dependent feature vector. P(X) and P(X|Y) denote the respective priori and posteriori probability of x and Y.

3.4.2 K-Nearest Neighbour Classifier

A straightforward technique called K-Nearest Neighbor categorizes incoming data or cases based on a similarity metric after storing all of the previous examples. A data point is often categorised using the classification of its neighbors. As a result, it forecasts value using a lazy learner. However, because of the relationship between classification time and data size, KNN is regarded as the slowest classifier. It makes predictions by using Euclidean distance to calculate the distance between the query point and the context in the samples. The texts are categorized using training samples and attributes. Finally, the KNN predicts if sentiments are good or negative.

3.4.3 Decision Tree Classifier

DT uses a hierarchical structure that can be used to handle classification and regression issues. DT distinguishes records with numerous features by comparing the property to the root and vertex of a tree. Positive or negative class labels are given to each terminal vertex. The presence or absence of at least one word is checked by the checker on the property. Up until there are the fewest amount of records, the tree is partitioned.

3.4.4 Logistic Regression Classifier

LR is typically used for supervised classification tasks. For a specific collection of features (or inputs), X, the target variable (or output), y, can only take discrete values in a classification issue. Regression modeling is used. The model creates a regression model to forecast the probability. It makes a prediction about whether a set of data belongs into the "1" category.Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function:

$$g(z) = 1/(1 + e^{-z})$$

3.5 Summary

In this section, we selected 10k reviews from the database and using the IMDB rating system, the reviews are divided into positive and negative. Firstly, we prepare the data where we cleanup the unused or confusing information which will not aid in the training. Then we use TF-IDF to determine the most crucial words. It is usually used to find keywords or to rapidly summarize articles. Afterwards, we use different classifier to

determine whether a movie is favorable to watch or not. The commonly used ML based model for text based sentiment analysis is presented in Fig.3.2. As illustrated the model consists of five main blocks along with few minor components integrated in the system. [12]

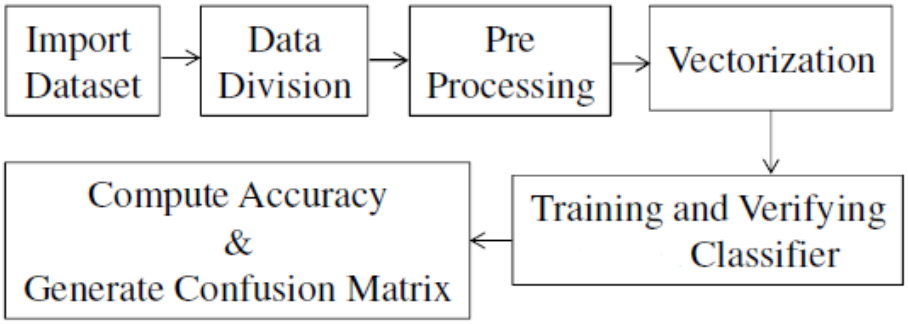

Figure 3.2. ML based model for text based sentiment analysis.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Performance Measure

Every pipeline for machine learning includes performance indicators. They tell us if we're making progress, and put a number on it. All machine learning models, whether they use SOTA methods like BERT or linear regression, need a metric to assess performance.
The performance indicators of every machine learning activity may be reduced to either Regression or Classification. For both issues, there are numerous measurements. Understanding how our model interprets our data is crucial. [13]

4.1.1 Accuracy

The classification accuracy is used to measure that how well the devised model is able to automatically identify the data. It is the percentage of labels that have been correctly classified. The mathematical formulation for accuracy is given below where TP, TN, FP, and FN respectively denote the true positives, true negatives, false positives, and false negatives in the predicted labels.

$$Accuracy = TP + TN/TP + FP + FN + TN$$

4.1.2 Recall

Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

$$Recall = TP/TP + FN$$

4.1.3 Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = TP/TP + FP$$

4.1.4 F1 Score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

$$F1Score = 2 * (Recall * Precision)/(Recall + Precision)$$

4.2 Result

In Naive Bayes the plot of confusion matrix is shown in Fig.4.1

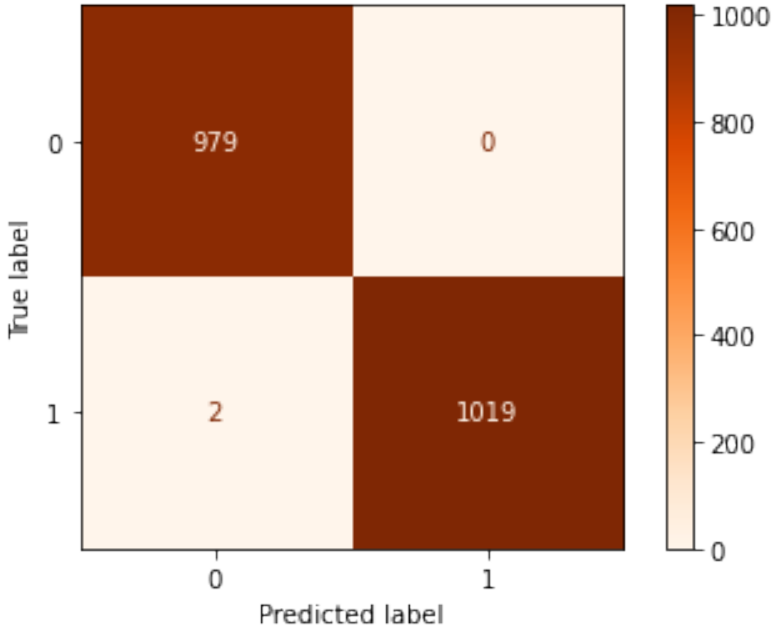

Figure 4.1. Confusion matrix plot for Naive Bayes Classifier.

In K-Nearest Neighbour Classifier the plot of confusion matrix is shown in Fig.4.2

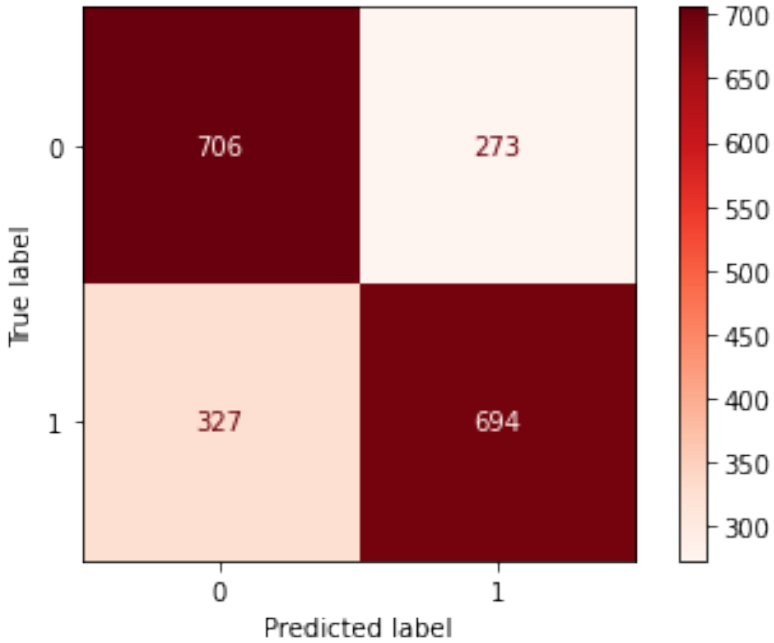

Figure 4.2. Confusion matrix plot for K-Nearest Neighbour Classifier.

In Decision Tree Classifier the plot of confusion matrix is shown in Fig.4.3

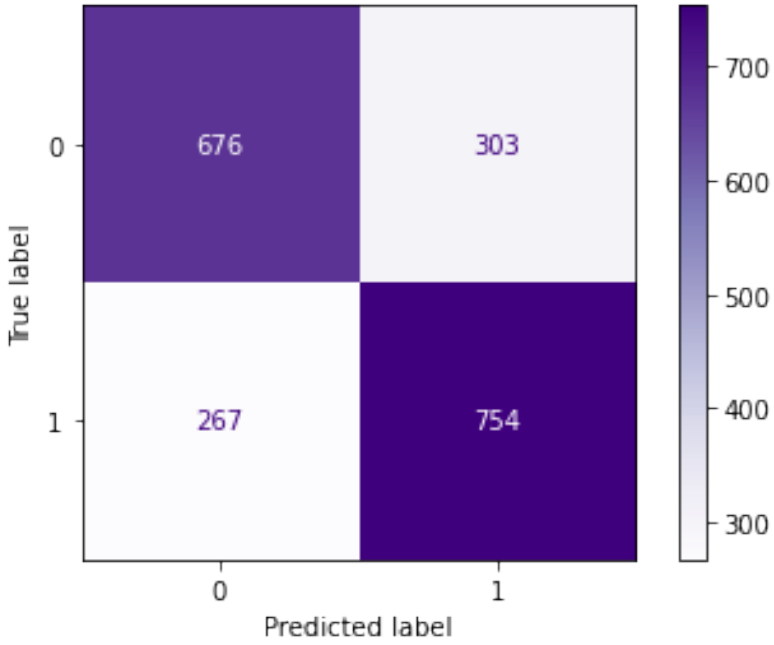

Figure 4.3. Confusion matrix plot for Decision Tree Classifier.

In Logistic Regression Classifier the plot of confusion matrix is shown in Fig.4.4

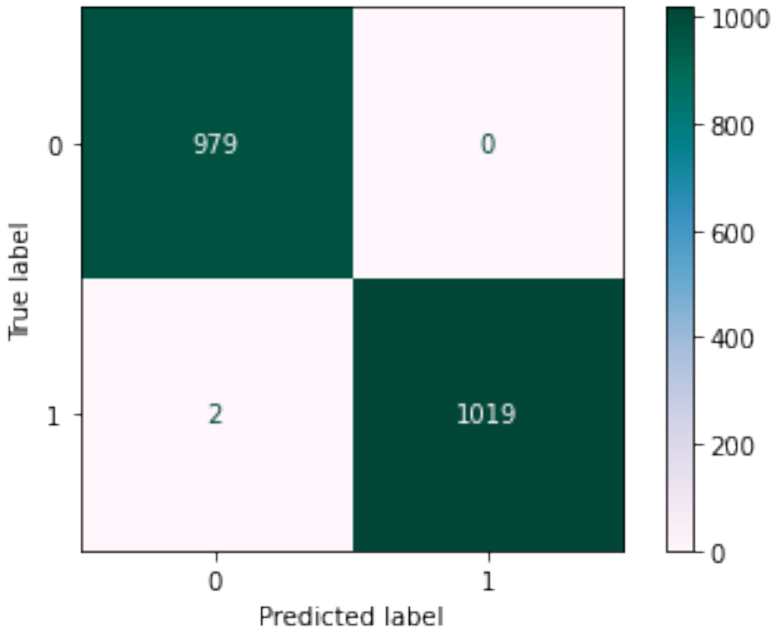

Figure 4.4. Confusion matrix plot for Logistic Regression Classifier.

Classifier	Reviews	Precision	Recall	F1-Score	Accuracy
Naive Bayes	positive negative	0.61 0.65	0.65 0.61	0.63 0.63	0.6295
k-Nearest Neighbor	positive negative	0.68 0.72	0.72 0.68	0.70 0.70	0.7
Decision Tree	positive negative	0.72 0.71	0.69 0.74	0.70 0.73	0.715
Logistic Regression	positive negative	0.67 0.86	0.67 0.86	0.67 0.86	0.8

Table 4.1. Generated result from all the classifiers.

The table-4.1 shows precision, recall, f1-score, and accuracy of all the classifiers for both positive and negative values. We can see that, Logistic Regression gives the accuracy of 0.8 percent which is better than other classifiers.

4.3 Summary

We have compared the result of the classification model based on their accuracy. We observed that Logistic Regression gave the best accuracy than the others. We went through the steps where we could understand about the accuracy, recall, precision, f1-score and matriculated the ways to calculate those. Afterwards, we generated all the plot for confusion matrix and also showed the generated result table of all the classifiers.

CHAPTER 5

CONCLUSION

In this report, we used natural language processing to find out the sentimental analysis of IMDB movie reviews. There are some limitations of this model:

- We didnot use bag of words or Word2Vec with gensim for vectorizing the dataset, so we donot know whether it will give better results than TF-IDF
- Our model cannot read the reviews which are in other languages. We could have used those libraries for definite languages to help our model read those
- Our dataset had 50k reviews in it, but we could not use whole of it because of hardware limitation so we shrinked it to 10k.

In future, we wish to conduct a similar study on different languages specially on Bangla. We also wish to experiment this dataset with different sentimental analysis methods in order to increase the accuracy of the results.

BIBLIOGRAPHY

[1] *Sentiment analysis of imdb movie reviews using long short-term memory*, Available online: click here (accessed on: 23 September 2022).

[2] *Learning word vectors for sentiment analysis*, Available online: click here (accessed on: 23 September 2022).

[3] *Movies reviews sentiment analysis and classification*, Available online: click here (accessed on: 23 September 2022).

[4] *Comparative study on sentiment analysis on imdb dataset*, Available online: click here (accessed on: 23 September 2022).

[5] *Sentiment analysis of movie reviews using machine learning techniques*, Available online: click here (accessed on: 23 September 2022).

[6] *Sentiment analysis of movie reviews a new feature-based heuristic for aspect-level sentiment classification*, Available online: click here (accessed on: 23 September 2022).

[7] *Sentiment analysis on movie reviews using naïve bayes classifier*, Available online: click here (accessed on: 23 September 2022).

[8] *Sentiment analysis of imdb movie reviews*, Available online: click here (accessed on: 23 September 2022).

[9] *Sentiment analysis for movies reviews dataset using deep learning models*, Available online: click here (accessed on: 25 September 2022).

[10] *Deep learning for sentiment analysis of movie reviews*, Available online: click here (accessed on: 25 September 2022).

[11] *Accuracy, precision, recall f1 score: Interpretation of performance measures*, Available online: click here (accessed on: 23 September 2022).

[12] *Sentiment analysis of imdb movie reviews using long short-term memory*, Available online: click here (accessed on: 23 September 2022).

[13] *Performance metrics in machine learning*, Available online: click here (accessed on: 23 September 2022).