Sentiment Analysis On Movie Rewiews

BY

Md. Shamsuzzaman

ID: 151-15-4861

Tapan Chandra Sarkar

ID: 151-15-4860

Konika Paul

ID: 151-15-4757

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Subroto Nag Pinku

Lecturer

Department of CSE, Daffodil International University

Co-Supervised By

Ms. Nishat Sultana

Lecturer

Department of CSE, Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

NOVEMBER 2018

APPROVAL

This Research titled **"Sentiment Analysis On Movie review"**, submitted by **Md. Shamsuzzaman, Tapan Chandra Sarkar and Konika Paul** to the Department of Computer Science & Engineering, **Daffodil International University**, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc in Computer Science & Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

Prof. Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Abul Hasnat Mohammad Saiful Islam — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin — **External Examiner**
Professor and Chairman
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this research has been done by us under the supervision of **Subroto Nag Pinku,** Lecturer and co-supervision of **Ms. Nishat Sultana** Lecturer, Department of CSE, Daffodil International University. We also declare that neither this research nor any part of this research has been submitted elsewhere for the award of any degree.

Supervised By:

Subroto Nag Pinku
Lecturer
Department of CSE
Daffodil International University

Co- Supervised by:

Ms. Nishat Sultana
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Md. Shamsuzzaman
ID: 151-15-4861
Department of CSE
Daffodil International University

Tapan Chandra Sarkar
ID: 151-15-4860
Department of CSE
Daffodil International University

Konika Paul
ID: 151-15-4757
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to Supervisor **Subroto Nag Pinku**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to the Almighty Allah and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Sentiment analysis indicates to the use of NLP (Natural Language Processing), text analysis and identify subjective information in source materials. The detection of Subjective information such a opinions, attributes, emotions or feelings what we are working on. In this Thesis we are working on IMDb movies review using text analysis finding sentiment analysis. At the very beginning we have used python built in libraries to execute the properties which we have used in our code. Then we collect the data set from Kaggle.com both train and test data. To built our project models we have used naïve bayes algorithm for probability measurement the sentiment. Logistic regression for portray information. Confusion matrix help us to find out precision., recall fi-score, support. For the input data we have used IMDb movie review comments and then we get the sentiments as an output.

TABLE OF CONTENTS

CONTENTS **PAGE**

Chapter 1
Introduction

1.1 Introduction

Expressing opinions and posting reviews about places visited or movies seen has become really popular nowadays. This has prompted the need to naturally comprehend this gigantic measure of information. The human language is complex therefore teaching a machine to analyze the various grammatical nuances, cultural variations, slang and misspellings that occur in reviews provided by users is a difficult process. Training a machine to see how setting can influence tone is much more troublesome. Progressions in machine learning and normal dialect preparing systems made it conceivable to examine client surveys and recognize the client's sentiments towards them. This methods of sentiment analysis are useful in a wide range of domains, such as business or politics. The purpose of the challenge proposed by Kaggle is to experiment different machine learning models for the task of sentiment analysis, using the Rotten tomatoes movie review corpus. More specifically, phrases from the dataset have to be classified in 5 categories, according to the intended tone of the user: negative, somewhat negative, neutral, somewhat positive, positive.

This paper represents the the final report of the project. In the third part we present the database gave on this test, trailed by the preprocessing systems which we connected on this database. From there on, we present the methods we have investigated and implemented so far. Finally we present our results and its interpretations.

1.2 Motivation

We did this research to see how many people like or dislike a movie. So many movies release in a year. But every movie is not good. There are so many good movies but how many movies are good that should be published. For this job collecting public opinion is the best process. We can collect public opinion by online or offline. Then we can gather the opinions and calculate the result by train our machines. By the result directors can see the result and can realize that what kind of movies people want. For entertaining the audiences they will eager to make people demanded movies.

1.3 Rationale of the Study

We use sentiment analysis to find an answer from public about their choice to watch a movie. We don't know normally What kind of movie, drama, character, vfx, movie circumstances and location people like. It is difficult to interview every people. So tough to gather opinions about a movie in offline. But it is easy to gain information about people's choice in online. It is the easiest thing to gain data about a movie that the movie is good or bad. To give people some good movies to improve their knowledge or to improve their mentality we can gather public opinions so that in the future we can make more good movies. A society's culture is one of the most important things to improve a society's children. If children can watch good movies their knowledge, mentality, strength will grow. If nudity spreads by movies children's mentality will destroy. So we need to gather public opinions for making good movies. For this reason we start a research on sentiment analysis on movie reviews. To make this analysis we use some algorithm like naïve bayes algorithm, SVM, confusion matrix and logistic regression.

1.4 Research Questions

We have gotIs it possible to find an answer from public by sentiment analysis?

1.5 Expected Output

In this project it has multiple categorical class labels. Vowpal Wabbit assumes labels to be positive integers, beginning from one. It could be identified by five labels namely very negative, negative, neutral, positive, very positive as [0,1, 2, 3, 4] respectively. When we take input then output will be 0 or 1 or 2 or 3 or 4. this integer number carry individual polarity.

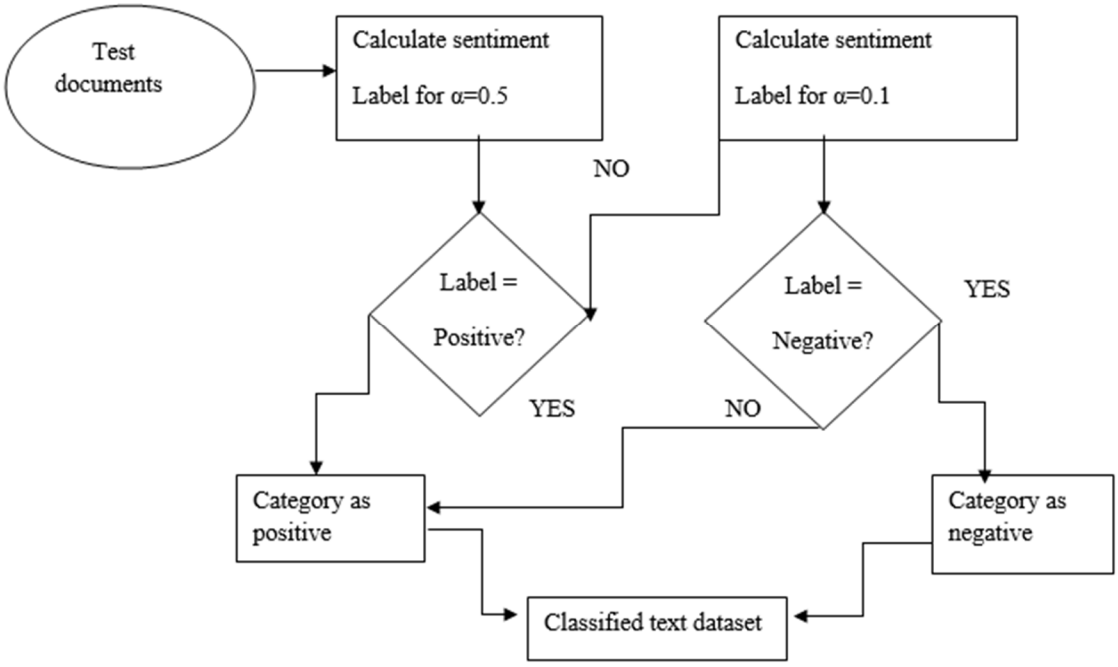

Figure 1.5: Flowchart of final classification of test dataset

1.5 Report Layout

- Chapter one have demonstrated an introduction to the project with its motivation, research questions, and expected outcome.
- Chapter two will have "Background" demonstrates introduction, related works, research summary, scope of the problem and challenges.
- Chapter three have Research Methodology.
- Chapter four will have Experimental Results and Discussion.
- Chapter five will have Summary and Conclusion.

Chapter 2

Background

2.1 Introduction

Online opinion become a most important and virtual currency with the conception of ratings, reviews, recommendation and other forms of online expression, for business that are looking to market their products. Automate is the process of filtering out the noise. Understanding the conversion. So that we can identify the relevant content and follow appropriate action Sentiment analysis is working on it. The main task in sentiment analysis is classifying the polarity of a given text, document, sentences. It point on whether the expressed opinion is positive, negative or neutral. Sometimes the emotional status go high such as "angry", "sad", "happy".

2.2 Related Works

Tirath Prasad Sahu, Sanjeev Ahuja worked on a study on feature selection & classification algorithms[1]. It was on sentiment analysis on movie reviews also. It has the 63% accuracy. V. K. Singh,R. Piryani,A. Uddin,P. Waila worked on movie reviews also[2]. Their

methodology was a new feature-based heuristic for aspect-level sentiment classification. Their accuracy was 75%. We read an another paper which was written by Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins[3]. They used text classification using string kernels. Their outcome was 57%. Lukasz Augustyniak, Tomasz Kajdanowicz, Przemyslaw , Kazienko ,Marcin Kulisiewicz Wlodzimierz Tuliglowicz worked on the methodology Lexicon Based vs. Classification[4]. The outcome was 60%. Rishanki Jain worked on Sentiment Analysis on YouTube Movie Trailer comments to determine the impact on Box-Office Earning[5]. His methodology was textual and classification. Deepa Anand,Deepan Naorem worked on Semi-supervised Aspect Based Sentiment Analysis for Movies Using Review Filtering[6]. The methodology was classification algorithms. Pimwadee Chaovalit and Lina Zhou worked on Movie Review Mining[7].Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, BeepaBose, Sweta Tiwari worked movie review by Naïve Bayes' and K-NN Classifier. Their outcome was 77%[8]. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher worked on movie reviews also and their methodology was Learning Word Vectors for Sentiment Analysis[9].

2.3 Research Summary

SL	Author	Methodology	Description	Outcome
1	Tirath Prasad Sahu, Sanjeev Ahuja	A study on feature selection & classification algorithms	Sentiment analysis of movie reviews	63%
2	V. K. Singh,R. Piryani,A. Uddin,P. Waila	A new feature-based heuristic for aspect-level sentiment	Sentiment analysis of movie reviews	75%

		classification		
3	Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins	Text classification using string kernels	Sentiment analysis of movie reviews	57%
4	Lukasz Augustyniak, Tomasz Kajdanowicz, Przemyslaw , Kazienko ,Marcin Kulisiewicz Wlodzimierz Tuliglowicz	Lexicon Based vs. Classification	An Approach to Sentiment Analysis of Movie Reviews	60%
5	Rishanki Jain	textual and classification	Sentiment Analysis on YouTube Movie Trailer comments to determine the impact on Box-Office Earning	57%
6	Deepa Anand,Deepan Naorem	A study on feature selection & classification algorithms	Semi-supervised Aspect Based Sentiment Analysis for Movies Using Review Filtering	62%
7	Pimwadee Chaovalit and Lina Zhou	A Comparison between Supervised and Unsupervised Classification	Movie Review Mining	61%

		Approaches		
8	Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, BeepaBose, Sweta Tiwari	Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier	Sentiment analysis of movie reviews	77%
9	Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher	Learning Word Vectors for Sentiment Analysis	Movie review analysis	63%

2.4 Scope of the Problem

In this project the main problem is collection of data set. Because when we search the train and test dataset it seems that this kinds of data set have nowhere and most of the time the text written non english format. lot bag of word found in the textual files.

2.5 Challenges

With accelerated evolution of the internet as websites, social networks, blogs, online portals, reviews, opinions, recommendations, ratings, and feedback are generated by writers. This writer generated sentiment content can be about books, people, hotels, products, research, events, etc. These sentiments turn out to be exceptionally useful for organizations, governments, and people. While this content meant to be useful, a bulk of this writer generated content require using the text mining techniques and sentiment analysis. But there are several challenges faced the sentiment analysis and evaluation process such as. These challenges become obstacles in analyzing the accurate meaning of sentiments and detecting the suitable sentiment polarity. Sentiment analysis is the practice

of applying Natural Language Processing and Text Analysis techniques to identify and extract subjective information from text.

Chapter 3

Research Methodology

3.1 Introduction

For this research we use Anaconda, NumPy, Jupyter and Matplotlib. We use Anaconda because it is best for data science and machine learning applications. To add large multi-dimensional arrays we use NumPy. We use Jupyter and Matplotlib also.

3.2 Research Subject and Instrumentation

Our research subject is sentiment analysis of IMDb movie reviews.

Anaconda

Anaconda is an open source. It is a free and open-source distribution of the Python and R programming languages. This is used for data science and machine learning applications (large-scale data processing, predictive analytics, scientific computing), that aims to streamline bundle administration and arrangement.

NumPy

NumPy is a library. It is used for the Python programming language to add support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. The ancestor of NumPy, Numeric. It was originally created by Jim Hugunin. It was the contributions from several other developers. In 2005, Travis Oliphant created NumPy by incorporating features of the competing Num array into Numeric, with extensive adjustments. NumPy is open-source software. This software has many contributors.

Jupyter

Project Jupyter is a nonprofit organization. This organization was made for creating open-source programming, open-measures, and administrations for intuitive figuring crosswise over many programming languages.

Matplotlib

matplotlib is a plotting library for the Python programming language and its numerical science augmentation NumPy. It gives a protest situated API to inserting plots into applications utilizing broadly useful GUI toolboxs like Tkinter, wxPython, Qt, or GTK+. There is likewise a procedural "pylab" interface dependent on a state machine (like OpenGL), intended to nearly look like that of MATLAB, however its utilization is discouraged. SciPy makes utilization of matplotlib.

3.3 Data Collection Procedure

We get the data from kaggle.com. This site contain huge datasets. The dataset has been divided into phraseId, SentenceId Phase, Sentiment. Every enties have aunique PhaseId, each sentence has SentenceId. Sentimental index describe the sentiment of the sentence. The train.tsv contains their phrases and associated sentence labels. The test.tsv contains just phases. Sentiment label to each phrase in text file should be assigned. The sentimental labels used in the dataset are-

0-Bad review

1-Somewhat bad

2-Average

3-Somewhat good

4-Very good

3.4 Statistical Analysis

We collected dataset from Kaggle . Then we trained the dataset the output of the dataset sentiments are-

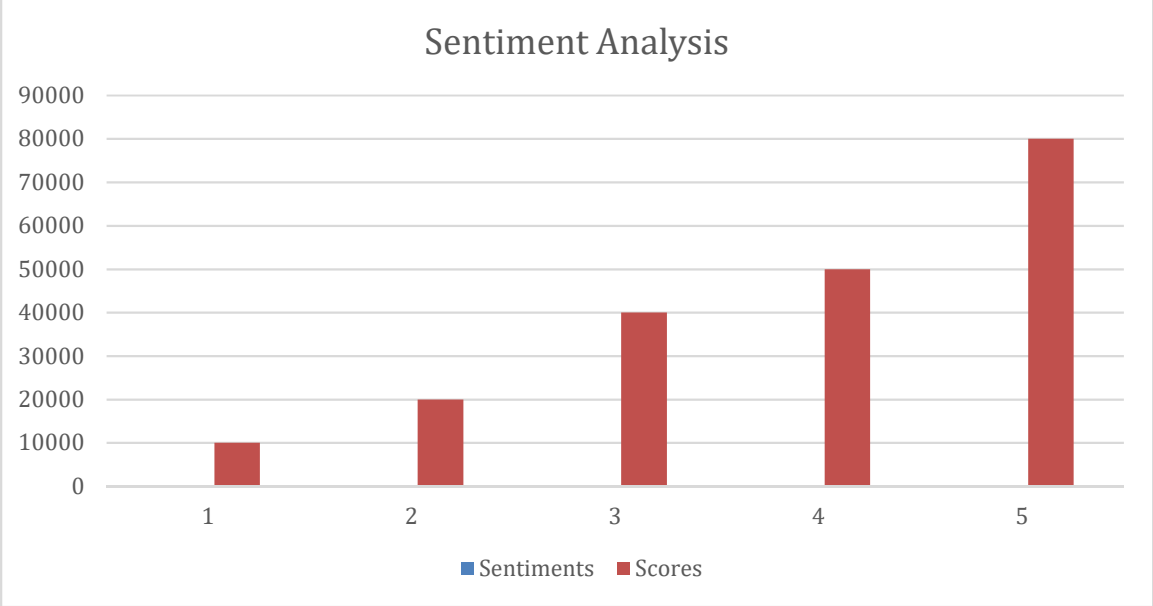

Figure 3.4.1: Output of the dataset sentiments

1. Randomly split movie reviews into 2 parts(75%25%).
2. Build Vectorizer Classifier pipeline (TfidfVectorizer)-

    -Eliminate rare and most frequent tokens.

    -Fit linear support classifier with relatively high frequency.

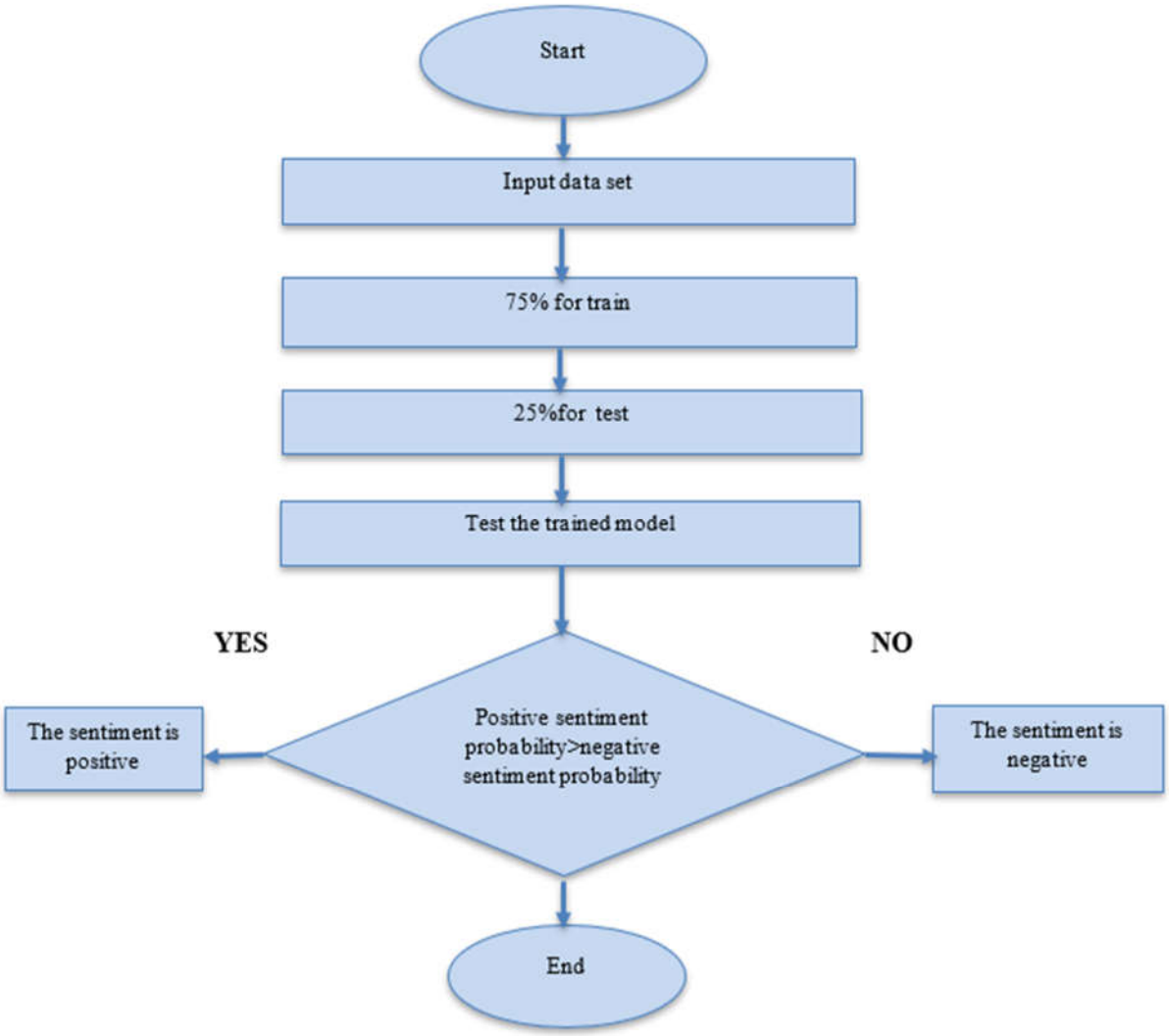

Figure 3.4.2: Flowchart of the working procedures

3.5 Implementation Requirements

We have applied some algorithms. They are Naïve Bayes algorithm, Logistic Regression, SVM, Confusion Matrix.

3.5.1 Naive Bayes Algorithm

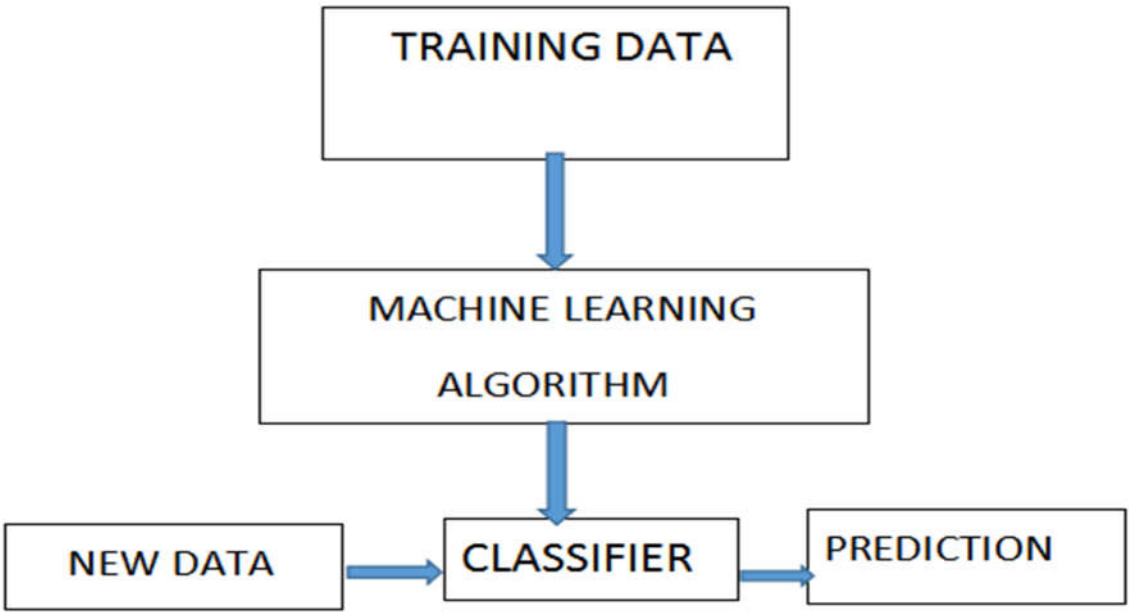

Figure 3.5.1: Naïve bayes algorithm process

Naive Bayes algorithm is an arrangement of directed learning calculations dependent on applying Bayes‟ hypothesis with the naive presumption of freedom between each match of highlights. Execution have been assessed utilizing holdout approach (60% - preparing, 40%- testing). This theorem will predict the result.

The Naive Bayes (NB) algorithm is an arrangement strategy dependent on Bayes' theorem with the presumption that all highlights are free of one another.

Bayes' theorem is represented by the following equation:

$$P(A|B) = \frac{P(A|B)P(B)}{P(A)}$$

Where A and B are features

- P(B|A) is the probability of B given A.
- P(A|B) is the probability of A given B.
- P(B) is the prior probability of B.
- P(A) is the prior probability of B.

The NB equation can be written as follows:

$$P(B|A)=P(a1|B)\ P(a2|B)\ \ldots\ldots P(an|B)\ P(B)$$

Where A = (a1, a2,… an) represents a vector of n features.

3.5.2 Logistic Regression

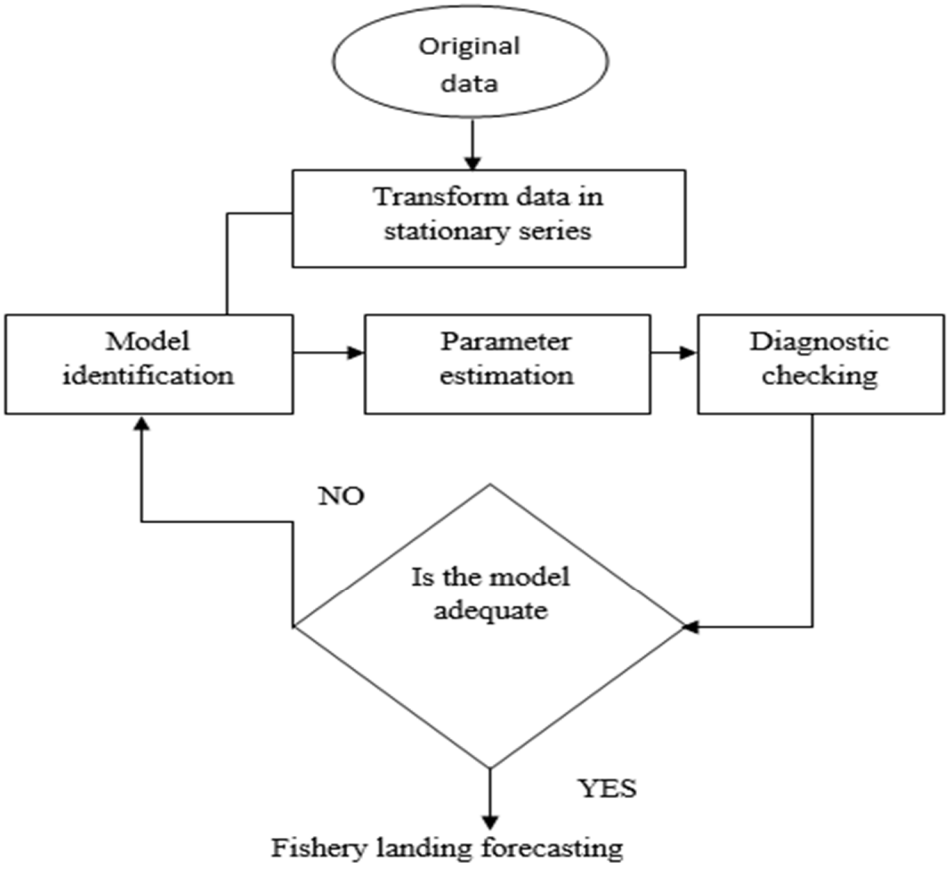

Figure 3.5.2: Logistic regression process

Logistic Regression is the suitable regression analysis to lead when the reliant variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Calculated regression is utilized to portray information and to clarify the connection between one ward paired variable and at least one ostensible, ordinal, interim or proportion level autonomous factors.

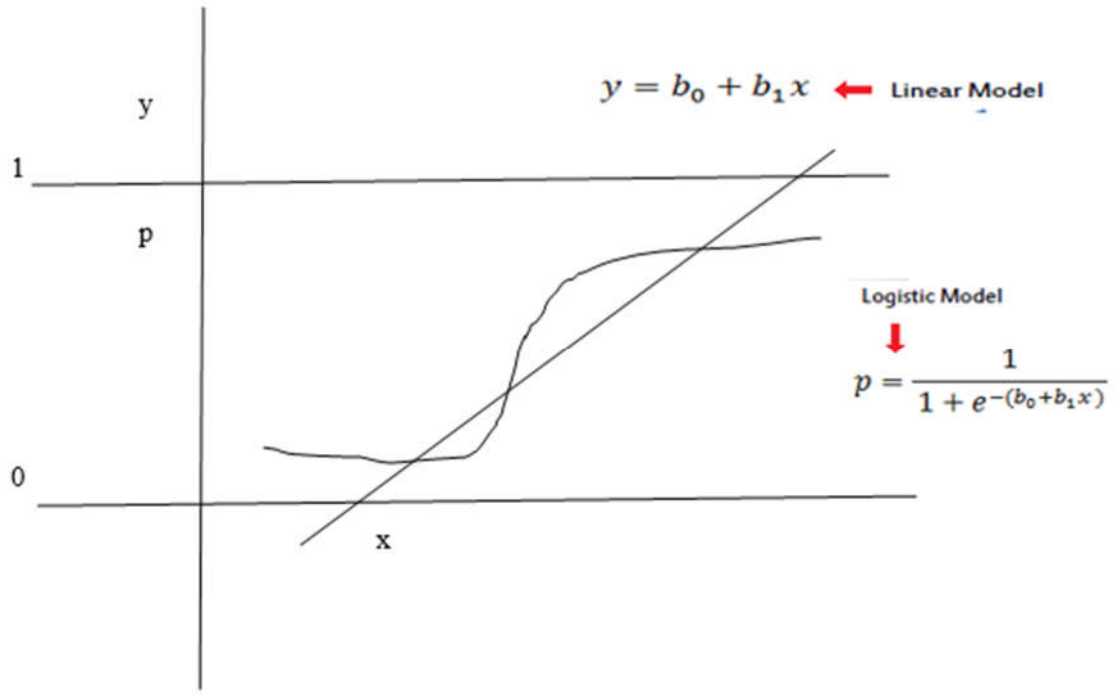

Figure 3.5.3: Logistic regression graph

3.5.3 SVM

Support vector machine is supervised learning model. This model analyzes data used for classification and regression analysis. Svm uses a subset of training points in the decision function called support vectors, so it is also memory efficient. It is a set of regulated learning techniques utilized for grouping, relapse and exception discovery. It attempts to discover a hyperplane that can viably partition the given preparing information into two sections. The significant preferred standpoint of help vector machines is viability in high dimensional spaces. Additionally it utilizes a subset of preparing focuses in the choice capacity called bolster vectors, so it is likewise memory proficient. The one disadvantage in SVM is when preparing information is profoundly lopsided, coming about model has a tendency to perform well on larger part information however perform terrible on minority information. In scikit library, distinctive sorts of portions, for example, straight, rbf and polynomial are given. It executes the multi class grouping utilizing one against one methodology.

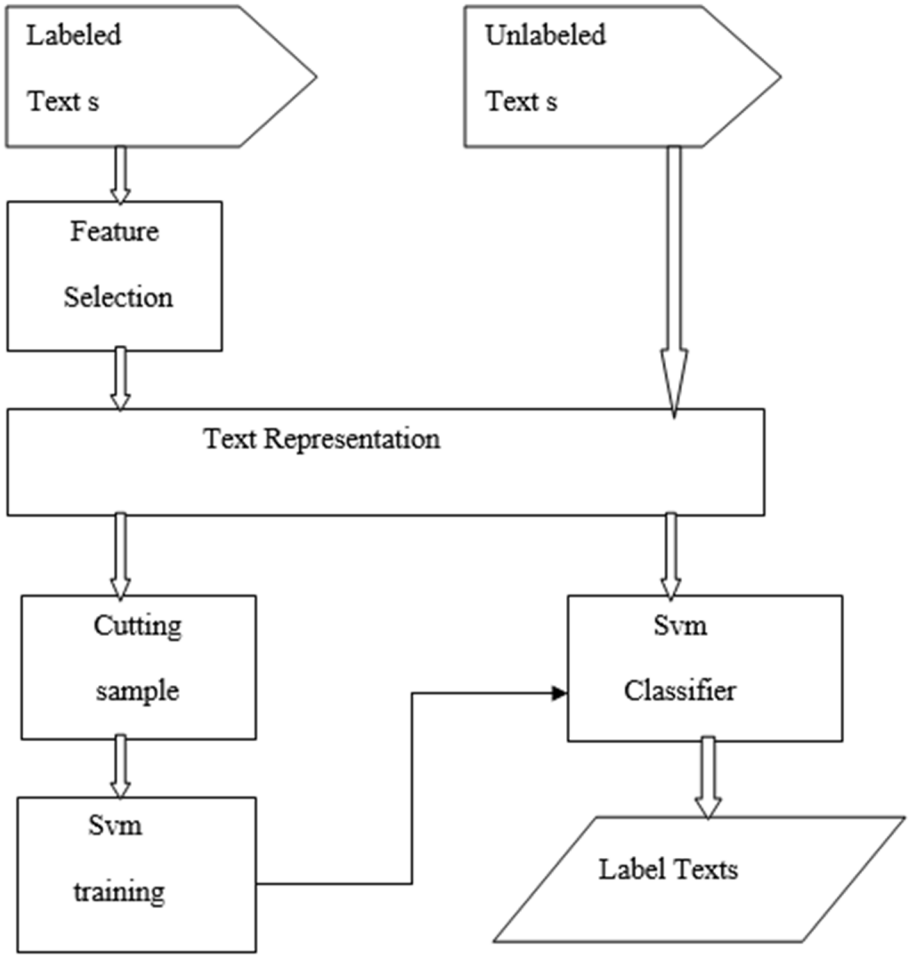

Figure 3.5.4: SVM process

3.5.4 Confusion Matrix

The Confusion Matrix is obtained by using Confusion Matrix() library function. Using matrix classifies a clear conception about positive negative or neutral values on a data set. It is basically a table that can be generated for a classifier on a binary data set and can be used to describe the performance of the classifier. It shows false positive false negative true positive true negative values.it just gives a ratio doing predictive values calculating from data set. Like I have a disease or I never have a disease. Here two types of values occur either he have a disease or he have not. there is a true positive or a false positive answer occur. Accuracy can be measured in parameter and it is calculated by number of correctly predicted reviews divide by total number of reviews present in the corpus.

Chapter 4

Experimental Results and Discussion

4.1 Introduction

First we take the train data and train the machine. Then we take the test data for test. After that we can input text.

4.2 Experimental Results

True Positive (TP): The sentiment is fully positive.

True Negative (TN): The sentiment is slightly positive.

False Positive (FP): The sentiment is slightly negative.

False Negative (FN): The sentiment is fully negative.

Precision: precision is the piece of related instances among the retrieved instances. high precision means that an algorithm returned substantially more relevant results than irrelevant ones.

$$precision = \frac{tp}{tp + fp}$$

Recall: Recall is the piece of relevant instances that have been retrieved over the total amount of relevant instances. High recall means that an algorithm returned most of the relevant result.

$$Recall = \frac{tp}{tp + fn}$$

F-measure: f-score is a measure of test's accuracy by considering both precision and recall. it is a harmonic average of precision and recall.

$$F - score = 2 * \frac{precision * recall}{precision + recall}$$

Accuracy: accuracy refers to the familiarity of the measured value to a known value.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

False Positive Rate: False positive rate refers that our proposed method predict the textual format is slightly positive. Calculate the false positive rate by the given equation:

$$F - score = 2 * \frac{precision * recall}{precision + recall}$$

Table i: Classification report

Number	precision	recall	Fi -score	support
0	0.44	0.30	0.36	1740
1	0.49	0.44	0.47	6854
2	0.69	0.78	0.73	19635
3	0.53	0.50	0.51	8384
4	0.49	0.33	0.39	2402
AVG/TOTAL	0.60	0.61	0.60	39015-

4.3 Descriptive Analysis

First we collect data from kaggle.com. Then we train and test the data. In our research 75% are train data and 25% are test data. We use some algorithms. The algorithms are SVM, confusion matrix, naive bayes and logistic regression. Naive bayes Algorithm is a quick, profoundly versatile calculation. Naive Bayes can be use for Binary and Multiclass grouping. We use this algorithm for prediction. We use Confusion matrix. Confusion matrix is a system for outlining the execution of a classification algorithm. Confusion matrix shows that the sentiment is positive, negative, true positive and true negative. We use SVM to analyze data used for classification and regression_analysis. Then we use logistic regression for predictive analysis.

When we put any new data then the algorithms classified the data and finally give us a predictive result.

4.4 Summary

At the very begining, we use some libraries for the further requirement of the code. Take the dataset from kaggle and read them both train and test. We also view them statictacally by using import matplotlib. pyplot as plt. Then split the train data into dependent and Independent Columns to find phrase and sentiments and their shape of both x shape and y shape.then Spliting the data into train/test by vect.fit_transform(X_train) and vect.transform(X_test).For building models we have used Naive Bayes algorithm.also set the confusion matrix into the building models. After that we do some parameter tuning on Logistic Regression. We also use SVM(Support Vector Method).At last to Predicting on input data sentiment we have used log. predict. Then we able to get the expected output.

Chapter 5

Summary, Conclusion, Recommendation and Implication for Future Research

5.1 Summary of the Study

In this work, we studied a wide range of NLP classification models. Our investigations consisted of main parts. we used the dataset provided by Kaggle and applied the bag of words, and skipgram word2vec models to represent words numerically. We then used several classifiers, including naive Bayes, SVM, and logistic regression to perform the classification task.

5.2 Conclusions

Therefore with the discussion which has been carried out it has been clear that proper preprocessing of data based on natural language processing approaches as well as incorporating already existing models in the domain of sentiment analysis altogether with appropriate classification process can improve the performance of the model for multiclass classification of movie reviews.

5.3 Recommendations

In this project it has multiple categorical class labels. Vowpal Wabbit assumes labels to be positive integers, beginning from one. It could be identified by five labels namely very negative, negative, neutral, positive, very positive as [0,1, 2, 3, 4] respectively. When we take input then output will be 0 or 1 or 2 or 3 or 4. this integer number carry individual polarity.

5.4 Implication for Further Study

Our next aim will be analysis the sentiments in our bangla language. If we can able to do that it will helpful or society to identify bad and good movies in our country. moreover, we can also analysis the sentiments the people who will participate the upcoming election in our country. For that we just have to train the political words and test them accurately.

References

[1] Tirath Prasad Sahu, Sanjeev Ahuja,"Sentiment analysis of movie reviews: A study on feature selection & classification algorithms",2016 International Conference on Microelectronics, Computing and Communications (MicroCom)

[2] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. Computer Science 4 (2014), 1188–1196.

[3] V. K. Singh,R. Piryani,A. Uddin,P. Waila,"Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification",2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)

[4] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In Empirical Methods in Natural Language Processing (EMNLP). 1532–1543. http://www.aclweb.org/anthology/ D14-1162

[5] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. Journal of Machine Learning Research 2, 3 (2002), 419–444.

[6]Lukasz Augustyniak Tomasz Kajdanowicz Przemyslaw Kazienko Marcin Kulisiewicz Wlodzimierz Tuliglowicz,"An Approach to Sentiment Analysis of Movie Reviews: Lexicon Based vs. Classification",2014 International Conference on Hybrid Artificial Intelligence Systems.

[7] Rishanki Jain,"Sentiment Analysis on YouTube Movie Trailer comments to determine the impact on Box-Office Earning 2015 International Conference on Microelectronics, Computing and Communications (MicroCom)

[8] Hieu Pham, Minh-Thang Luong, and Christopher Manning. 2015. Learning Distributed Representations for Multilingual Text Sequences. In the Workshop on Vector Space Modeling for Natural Language Processing. 88–94.

[9] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. Computer Science 4 (2014), 1188–1196.

[10] Deepa Anand,Deepan Naorem,"Semi-supervised Aspect Based Sentiment Analysis for Movies Using Review Filtering",2013 International Conference on Hybrid Artificial Intelligence Systems.

[11] Thorsten Joachims. 1997. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In Fourteenth International Conference on MachineLearning. 143–151.

[12] Shravan Vishwanathan," Sentiment analysis on movies reviews",3 rd IRF International Conference, 10th May-2014, Goa, India, ISBN: 978-93-84209-15-5.

[13] Pimwadee Chaovalit and Lina Zhou. 2005. Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches. In Hawaii International Conference on System Sciences. 112c.

[14] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002. Text classification using string kernels. Journal of Machine Learning Research 2, 3 (2002), 419–444.

[15] Asiri Wijesinghe,"Sentiment Analysis on Movie Reviews", rd IRF International Conference, 10th May-2014, Goa, India.

[16] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human

Language Technologies. Association for Computational Linguistics,Portland, Oregon, USA, 142–150. http://www.aclweb.org/anthology/P11-1015

[17] Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, BeepaBose, Sweta Tiwari, “Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier”, International Journalof Information Engineering and Electronic Business(IJIEEB),Vol.8, No.4, pp.54-62, 2016. DOI: 10.5815/ijieeb.2016.04.