COMPARATIVE SENTIMENT ANALYSIS USING DIFFERENCE TYPES OF MACHINE LEARNING ALGORITHM

BY

RAKIB HOSSAIN
ID: 161-15-6802

AND

FOWJAEL AHAMED
ID: 161-15-7045

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

Supervised By

Ms. Umama Dewan
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

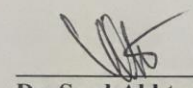
DHAKA, BANGLADESH

DECEMBER 2019

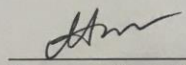
APPROVAL

This Project/internship titled "**Comparative sentiment analysis using difference types of machine learning algorithm**", submitted by Rakib Hossain, ID No: 161-15-6802 & Fowjael Ahamed, ID No: 161-15-7045 to the Department of Computer Science and 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 and Engineering and approved as to its style and contents. The presentation has been held on 05/12/19.

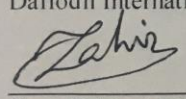
BOARD OF EXAMINERS



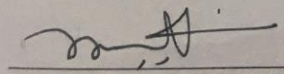
Dr. Syed Akhter Hossain — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Nazmun Nessa Moon — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Gazi Zahirul Islam — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

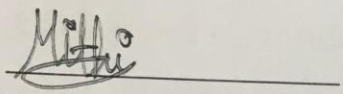


Dr. Mohammad Shorif Uddin — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

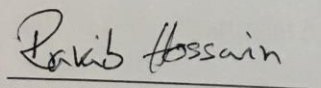
We hereby declare that, this project has been done by us under the supervision of **Ms. Umama Dewan, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

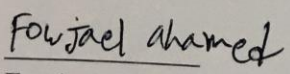


Ms. Umama Dewan
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Rakib Hossain
ID: 161-15-6802
Department of CSE
Daffodil International University



Fowjael Ahamed
ID: 161-15-7045
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 **Ms. Umama Dewan**, **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. Her 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 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

The company is becoming digital-focused in today's world company. Companies are selling their goods and searching for input from customers. It becomes difficult to say that the product is good or not based on their feedback when all the user writes their review about that is an item. That's where it comes to deep learning. By using this, we can derive thoughts or emotions from the consumer's written text. This is a study of emotion. It may determines the review's emotional status. Our plan senses views from analysis by the user whether they are good or bad. We use algorithms such as SVM, Naive Bayes, and some approaches. We use the algorithm of Naive Bayes because we want to learn how often in the text words occur. And then we use SVM to define positive or negative terms. For our researching purpose, we use the Amazon consumer review data set, which was available online. Some methods that we are using for preprocessing and cleaned the document where just words are left. We trained our model so well with twenty-four thousand data. So, it will give us the best accuracy and we make this model with the best algorithm and after that, it gives the accuracy of 98.39%. This project will help us in real life when we are having trouble with product reviews. Our machine will help us to determine which review is good and which review is bad and make a category of a positive and negative review and saves our time.

Keyword: Naïve Bayes, SVM, KNN, Polarity, Sentiment, Positive, Negative, Word, Paragraph, Accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Technology is the most important thing in today's world. Every human depend on this more and more. So the huge amount of data are created every moment. This data are used to develop the product by pre-processing. Opinions are the most important data for improved the product and it is the big research data for the recent world. It is very important to classify them like negative or positive. As an example "It's not suitable for him", it's must be negative type opinions. But the positive type of word 'suitable' will give a positivity. But the word 'not', are full changes the meaning of the sentence.so, it proves that we can't decide a sentence is a positive or negative basis on some keywords. It is important that a word is used after or before the keyword must be the consideration when made meaning of the sentence.

Sentiment Analysis is one of the Natural Language Processing (NLP) and Machine Learning (ML) applications that is widely studied. With the introduction of Web 2.0, this area has grown tremendously. The Internet has provided people with a platform to express their thoughts, emotions and feelings about goods, events, and life in general. Sentiment Analysis focuses on the recognition of whether a given piece of text is positive or objective, and whether it is constructive or negative.

The recent trends in the techniques of sentiment analysis have advanced towards the development of generative models capable of capturing complex contextual phenomena. By comparison, the emphasis is turning towards unsupervised approaches that use the power of co-occurrence to solve the problem due to the lack of annotated information. Since the internet has an enormous amount of opinionated information in the form of blogs, reviews, etc., the unmonitored approaches are flourishing.

It is very elaborate type problem to divine the sentiments given the text, but this problem solved using by Naive Bayes classifier, K neighbors classifiers and support vector machines. In this paper, I represent some classifier methodology like SVM and Sequential model, feature selection, words emphasizing and effective negation handling which is improved the accuracy of the result in sentiment analysis.

1.2 Motivation

Consumers and producers respect the "opinion of the consumer" about products and services. Therefore, both industry and academia have made considerable efforts to examine sentiments.

1) The Consumer's Perspective

When making a decision, having the opinion of the people around us is really important to us. Once upon a time, this community was small, with some trusted friends and family members. But we now see people expressing their opinions in blogs and forums with the advent of the Internet. These are now frequently read by people who are looking for an opinion on a particular entity (product, film, etc.). Therefore, on the Internet there are plenty of opinions available.

From the point of view of users, it is very important to gather feedback on a particular entity. It is difficult for consumers to try to go through such a large amount of information just by the sheer volume of this data to grasp the general opinion. Therefore, the need for a framework that separates good reviews from bad reviews. Furthermore, labeling these documents with their feelings would provide the readers with a succinct summary of an entity's general opinion.

2) The Societies Perspective

Recently, the Internet has caused some incidents that have impacted the government. Social networks are used to bring people together so that mass meetings can be coordinated and opposed to injustice. The social networks on the darker side are being used to insinuate people against an ethnic group or class of people, resulting in a significant loss of life. Therefore, Sentiment Analysis systems are required that can recognize and curtail these phenomena if necessary.

3) The Producer's Perspective

Consumers have at their fingertips, with the proliferation of Web 2.0 sites such as blogs, discussion forums, etc., a platform to share their brand experiences and opinions, positive or negative about any product or service. According to Pang and Lee (2008), these customer voices can have a huge influence in influencing the opinions of other customers and, eventually, their product loyalties, their buying decisions and their own brand advocacy.

Because customers have started to use the power of the Internet to expand their horizons, there has been a rise in review sites and blogs where users can perceive the advantages and faults of a product or service. Thus these views shape the product or service's future. Sellers need a program that can detect customer feedback trends and use them to enhance their product or service as well as predict potential requirements.

1.3 Rational of the study

In today's environment where we are justifiably suffering from data complexity (although this does not mean better or deeper insights), companies may have received mountains of customer feedback; but it is still difficult for mere humans to evaluate it manually without any kind of mistake or prejudice. Using large volumes of text material, sentiment analysis is useful to quickly gain insights. Besides the use case of customer feedback analysis,

which we touched on above, here are two more examples where sentiment analysis may be useful.

How could this information be useful when we look at the customer experience? If we want to reduce customer churn, we can use sentiment analysis to focus on the comments where the feeling is strongly negative. We can also look at customer feedback that has a strong positive feeling and figure out why these customers love us and simply focus on what we can do as a company to increase the number of our promoters.
With all the above examples, we can see the use of sentiment analysis by taking a source of text information with a broad context range and then gaging the text's polarity. We are researching in this topics for making business world better and make their customer better.

1.4 Research Questions

- How can this comparison help other researcher?
- How can this research help the business company?
- What is the future work of this comparative research project?
- How can general people can use this project in their daily life?
- Does anyone research about comparative research on sentiment analysis using different algorithm before?

1.5 Expected Output

We choose this project to make a system that decide which is best for the consumer and which is wrong by using review data and it can collect data and analyses them and predict which good product is and suggest them to user. In this project we use a comparison between four different models based on natural language processing and find out which model is the best for this kind of dataset and compare them. So that, we can use this model in our future work and make that kind of system which helps the user and companies to make profit for their company and attract the best customer for their product.

So, finally we decided to find out the best model for our project using python 3 for sentiment analysis and Natural Language processing for the dataset where we collect data from amazon and it contains the user's review and ratings.

1.6 Report Layout

This report contains a total of 5 chapters. In Chapter 1, the focus behind the thesis is on the introduction, motivation and goals. Chapter 1 is made up of six parts. 1.1, 1.2, 1.3, 1.4, 1.5 and 1.6 cover the introduction, inspiration, study logic, research problems, planned performance and document format. Chapter 2, consists of background information related to the thesis and is divided into 5 sections. Section 2.1, 2.2, 2.3, 2.4 and 2.5, consists of introduction, related work, study overview, problem context and individual challenges.

Chapter 3 describes the details covering 5 aspects of our study experiment. Sections 3.1, 3.2, 3.3, 3.4 and 3.5 describe each recognition phase, i.e. Introduction, subject of research and instrumentation, procedure for data collection, propose work follow model and implementation. The experimental results and discussion of that result were mentioned in Chapter 4. It's got four portions. Section 4.1, 4.2, 4.3 and 4.4 consists of introduction, findings of experiments, concise description and overview. Basically, Chapter 5 includes 4 parts with overview, conclusion, suggestion and implications for future research

CHAPTER 2

BACKGROUND

2.1 Introduction

The method of measuring and categorizing opinions expressed in a piece of text, in general to decide whether the attitude of the author towards a particular subject, item, etc. is positive, negative or neutral. Since it has many practical applications, sentiment analysis is currently a subject of great interest and growth. Organizations use sentiment analysis to continuously evaluate survey responses, product reviews, comments on social media and the like in order to gain valuable insights into their brands, products and services. There are many forms and varieties of feeling analysis and SA methods ranging from systems focusing on polarity (positive, negative, neutral) to systems detecting feelings and emotions (angry, happy, sad, etc) and recognizing desires (e.g. interested v. not interested). We're going to cover the most important in the following section.

Sentiment analysis in the field of technical communication is a rapidly growing subject. Increasing social media, online retail, and personal blogs and magazines, knowing where public sentiment is leaning, has transformed sentiment analysis into a rapid evolution that can transform into useful abilities.

Hybrid methods principle is very intuitive: simply combine the best of both worlds, rule-based and automatic. Normally, the methods can boost reliability and accuracy by combining all approaches.

Like rule-based systems, automated methods rely not on manually designed rules, but on techniques of machine learning. Typically the role of sentiment analysis is modeled as a classification issue where a classifier is fed with a text and returns the corresponding type, e.g. positive, negative or neutral.

The sentiment analysis applications is infinite. More and more we see it being used to track customer reviews, survey responses, rivals, etc. in social media tracking. Nevertheless, in business analytics and cases where text needs to be evaluated, it is also realistic for use.

2.2 Related Work

Thousand of data analyzer all over the world are trying to pursuit to increase the accuracy of sentiment analysis. Paul Ferguson team pursuit analysis to improve the accuracy of sentiment analysis based on paragraph level in 2009(6). In 2009, Yelena Mejova working with emotionally-charged text based on sentiment. Also working this field to develop sentiment accuracy, Christos Livas, Konstantina Delli and Nikolaos Pandis Invisalign patient testimonials on YouTube as well as the sentiment of the comments. Francesc Alías and Alexandre Trilla working with speech based sentiment in 2013(7). Aspect-based review analysis was done by Devina Ekawati and Masayu Leylia Khodra in 2017(8).

In the author of (1), Nurulhuda Zainuddin and his team worked sentiment analyzing using SVM. They worked using benchmark dataset to train the classifiers. They used to different weighting scheme and N-grams to extract the classical feature. They using square feature selection for improving the classification accuracy.

In(2) the author represent a model for classifying the movie reviews. They used machine learning to find out the difference between polarity classification and subjectivity detection and proposed a method for text-categorization. Show that the Naïve Bayes algorithm is more effective to show the result using subjectivity detection to shorter reviews.

In(3) The author proposed a model for classifying multiple language web forum reviews. To improve the performance the author used entropy weighted algorithm. They used SVM to get the to get higher performance with feature selection methods with high accuracy more than 90%.

In(4) The author proposed a method produce some feature automatically in a movie review analysis system using a multi knowledge-based approach. They combined the wordnet, statical analysis, and knowledge of movies. Their model was so effective it's proved that final output.

In author(5) S.M. Mazharul Hoque Chowdhury and his team approach a method to analyze a text in the paragraph level.The implementation of this method using a bag of words and priority based on the lexical analysis.

Nadeem Akhtar and his team(9) are reviewing the hotel feedback and providing information that may miss scores. The comments and metadata were crawled from the website and grouped according to some of the common aspects into predefined categories. Then the subject modeling technique (LDA) is applied to define hidden information and features, accompanied by an evaluation of feelings on confidential phrases and summarization.

Omar Raghib, Eshita Sharma, Tameem Ahmad and Faisal Alam (10) discuss the series of steps to be followed in this paper to analyze the speech signal for the recognition of their emotions highlighting some of the best available techniques for each step at present. During the transmission, distractions such as background noises make speech recognition complicated and difficult. A method for performing the classification of speech for different emotions is provided in this paper.

Vanshika Varshney, Aman Varshney and their team(11) are exploring the techniques of different machine learning algorithms to deduce a user's personality from their social media activities. Using three algorithms to compare the results of these three algorithms, namely SVM, KNN and MNB. Eventually, the author provides all three algorithms with a cumulative performance.

Palak Bansal, Somya, Nazar Kamaal, Shreya Govil, Tameem Ahmad (12) research is an effort to summarize the product reviews of consumers in a more usable and concise version that can help other users make their decisions. Web reviews are crawled of product, the first detection of product features will be performed every time after extraction and therefore polarity will be identified, i.e. either a review is positive review or a negative review. The description of all product features will be produced after the calculations.

Istuti Singh, Anil Kumar Sahu(13) worked on this article, a study to examine the behavior of stone columns used in various building types, such as oil storage tanks, embankments, houses, etc. The effect of stone columns without encased and enclosed on several construction forms is being examined. Also checked the effect of different diameters with different depths in the soil. Various types of geosynthetics are used for the enclosure to enhance the performance. A variety of mathematical and physical methods were performed to predict the settlement of foundations reinforced with stone base. In this article, there are also several theories from past to present that help in understanding stone column enhancements to improve soft soils. Physical simulation has a major role in the development of geotechnical properties.

So, we proposed a method which is given better accuracy other than analysis.

2.3 Research Summary

In this paper, we have compared different algorithms using a common dataset which is collected from Amazon. This dataset is contained with many user review, ratings, username. We use this data set for detecting sentiment of each review and try to differentiate between four models such as KNN, Navie Bayes, Gradient Boosting, Support Vector Machines. We use those models to find out the accuracy of the sentiment that are given by them and differentiate between them. Many people are using single model for their research for finding the best accuracy.

We try to compare between them and find out what is the best algorithm for this kind of dataset and we try to make machine to understand what is positive review and what is negative review and what is neutral and by the basis of this computer can filter the best possible product for the consumer and make the best decision and it won't need any kind of human assistant. This is the main reason for our research and make the other researcher to understand the meaning of this human sentiment and how useful machine can be and how useful model can be based on same dataset. We divided the dataset between two parts such as train and test and then we use different kind of model to find out the best accuracy to find out which review is positive and which review is negative and which is neutral. And this is the whole scenario of our research in a single picture.

2.4 Scope of the Problem

Study of emotions is already progressing from general (positive, negative or neutral) to a much more complex and deeper definition of granularity. So, there is a growing demand for sentiment analysis on both the research and business side. Researcher are working on their accuracy of their algorithm and the development of their model for making the best accuracy and make that model best for that problem. In our project, we decide that a comparison between some models which gives us best accuracy using the same dataset for those four model. And we try to differentiate those models and which is best for this kind of dataset. We take some reviews from amazon and make a dataset for this research and use this dataset for this project. We want to try this because we want to develop a system where system can decide which is best for the consumer and which is bad by using consumer's review from that site. Here we use four model to make this. By using this research commercial company can change their products and make it better for the consumer and they know about their consumer's satisfaction by using this kind of research. They will be able to change their marketing policy to improve their benefits and attract new customers. And we want to contribute more to this kind of field by developing our work in future.

2.5 Challenge

With the evolution of modern technology, mankind uses modern technology to decide their feelings to other human being and sometimes writers make it so simple and express their feelings to others and by that texts we can't decide that writer is happy or sad because text can't express their feelings because those aren't alive and human can tell who is sad or who is happy by hearing their voice tone but for the text we can't decide that whether it is positive or negative. It is so difficult to decide that. It is the biggest challenges of sentiment analysis. These challenges end up boundaries in examining the accurate meaning of sentiments and detecting the appropriate sentiment polarity. Sentiment analysis is the

exercise of making use of natural language processing and text evaluation techniques to perceive and extract subjective data from text.

It is a very popular topic nowadays. When its research using machine learning, it's faced some issues analyzing the text. An example an opinion gives a positive review in one nation but it's different from the other nation. Here, some common issues in sentiment analysis:

- Interrogative sentence gives us incorrect score because it's didn't identify properly which is a positive or negative opinion.
- Some opinions are didn't identified. Because it's not used any sentimental word.
- Some opinions can be spam when analyzing reviews and it provides negative score although it's a positive review.
- Sarcastic opinions are a challenging part of the analysis to be handled. Some of the sarcastic opinions give us negative meaning but it looks like a positive review. That's the case it provides a different result. Although a maximum number of the sarcastic sentence are positive.
- Different types of language give different meaning in a single word. It's made hesitation gives accurate output.
- Different types of language use in a single language, it's doesn't healthy for accurate accuracy.

Suppose, "this is good" and "is this good?" are two sentences and we have to find which text express which opinion. If we remove all the space, punctuation, and auxiliary word then we see that those two sentence has same word and it is 'good'. By using this word we can decide that this two sentence express positive opinion but is it? No, we see that second sentence don't express same opinion it express opposite of the first sentence. So we can see that sentiment of some sentence of same words aren't the same. And that's why it is difficult to understand the meaning of the sentence because emotion is too much complex.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

We proposed a method for sentiment analysis before we need to a data set which used for analysis.so, Takes in a string of text input for analyzing which are collaborating a sentence or paragraphs. The text or reviews are loaded into the system before pre-processing. When pre-process the data set that's time to remove all the white space with punctuation, because it's not needed us for the sentiment. We also remove all the stop words before splitting the sentence into word, after that all the words are converted into lower case context. Finally, all the data return for text processing.

Before text processing, we need to separate the positive word, negative word and also objective. Sentiment analysis phase identification the available sentiment word. After identification and separation word we should make the label of data set, positive is one (1) and negative is zero (0). Next, we vectorize our input variable using count vectorizer function which returns a vector array. Finally modified the data set compressed spares row format using transform function ().

3.2 Research Subject and Instrumentation

Our research topic is "Comparative Sentiment Analysis Using Difference Types of Machine Learning Algorithm." It is the field of Natural Language processing system.

Up to now we have discussed the theoretical concepts and methods. Now a list of requirements of instruments are given below-

Hardware and Software instruments which we use -

- 6^{th} generation core i3-4160, 3.60 GHz with 8 GB RAM.
- Intel(R) HD Graphics 4400
- 1 TB HDD

Developing Tools –

- Windows 10 version 1903
- Python 3
- Numpy ,Pandas ,Sklearn ,Seaborn ,NLTK ,Matplotlib

3.3 Data Collection Procedure

It's an important part in this research where we applied some method before data training and testing called data preprocessing. Our model has applied some procedure for implementation. All the procedures have mentioned during this section. The total information assortment method has been divided into 4 totally different states. All those states are delineated in figure three 3.3

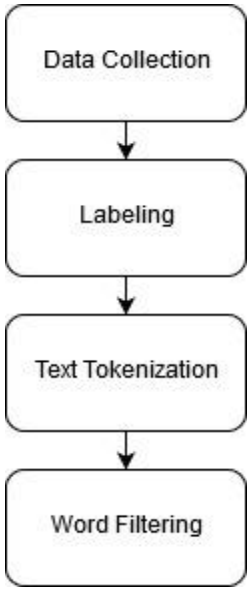


Figure 3.1: Flow of working procedure

1) Data

We collect the total 24178 reviews from amazon product review dataset and we transform the string into a meaning full vocabulary which are detect a review is positive or negative.

We get total 16954 meaning full word for classifying our data set. This are the following sample head of data set.

	Rating	Review	reviews.title	reviews.username
0	5.0	This product so far has not disappointed. My c...	Kindle	Adapter
1	5.0	great for beginner or experienced person. Boug...	very fast	truman
2	5.0	Inexpensive tablet for him to use and learn on...	Beginner tablet for our 9 year old son.	DaveZ
3	4.0	I've had my Fire HD 8 two weeks now and I love...	Good!!!	Shacks
4	5.0	I bought this for my grand daughter when she c...	Fantastic Tablet for kids	explore42

Figure 3.2: Collected Dataset

2) Text Tokenization

Text tokenization is a way to convert a sentence into the separate section. In this methodology, we convert a sentence into words where different types of punctuation are included. Every section is separate consider the white space in a sentence. When it's getting a full stop then consider the line reaches the endpoint. An example: "If you want something new, you have to stop doing something old.". In this example tokenize the whole sentence like 'If',' you',' want',' something',' new','(,)', 'you',' have',' to',' stop',' doing',' something',' old','(.)' where ',' and '.' Are punctuation and '.' The dot indicates the endpoint of this line. These whole things we complete using this NLTK method which tokenizes the many languages with English. In NLTK, Takes a string in a text and perform the following this task like remove all punctuation, remove all stop words and finally return the clean text into list words. We are using those below codes for tokenization.

```
tokens = X[0].split()
print(tokens)
```

```
['This', 'product', 'so', 'far', 'has', 'not', 'disappointed.', 'My', 'children', 'love', 'to', 'use', 'it', 'and', 'I', 'lik
e', 'the', 'ability', 'to', 'monitor', 'control', 'what', 'content', 'they', 'see', 'with', 'ease.']
```

Figure 3.3: Tokenization process

3) Word Filtering

After tokenization, unexpected word, number, and punctuation are removed that will not be effective in classification. So, at first, removed the punctuation and numbers which are not to require in classification. After remove that finally unwanted word are removed which didn't detect the positivity or negativity in a sentence. If we consider the previous example then we removed the all unauthorized object and get the filtering word are ' want', ' stop', ' doing', ' old', ' new', ' something'. NLTK is performing the following word filtering task.

```
sample_text = "Hey there! This is a sample review, which happens to contain punctuations."
print(text_process(sample_text))
```

```
['Hey', 'sample', 'review', 'happens', 'contain', 'punctuations']
```

Figure 3.4: Word filtering process

3.4 Proposed Model

A model is proposed when offered and an argument the element of the data set model and also discussed. We have proposed a model which get the input in text set. This text set has some opinion as review or comment which used in the analysis for taking sentiment. Two major layers in our proposed model one are data processing and another one is analyzing sentiment. Data processing layer discuss the collection of data, pre-processing data and mad prefer for sentiment analysis. Another layer processes the data for the train and testing in a method and gets the expected output. We will be discussed it's briefly in the following section.

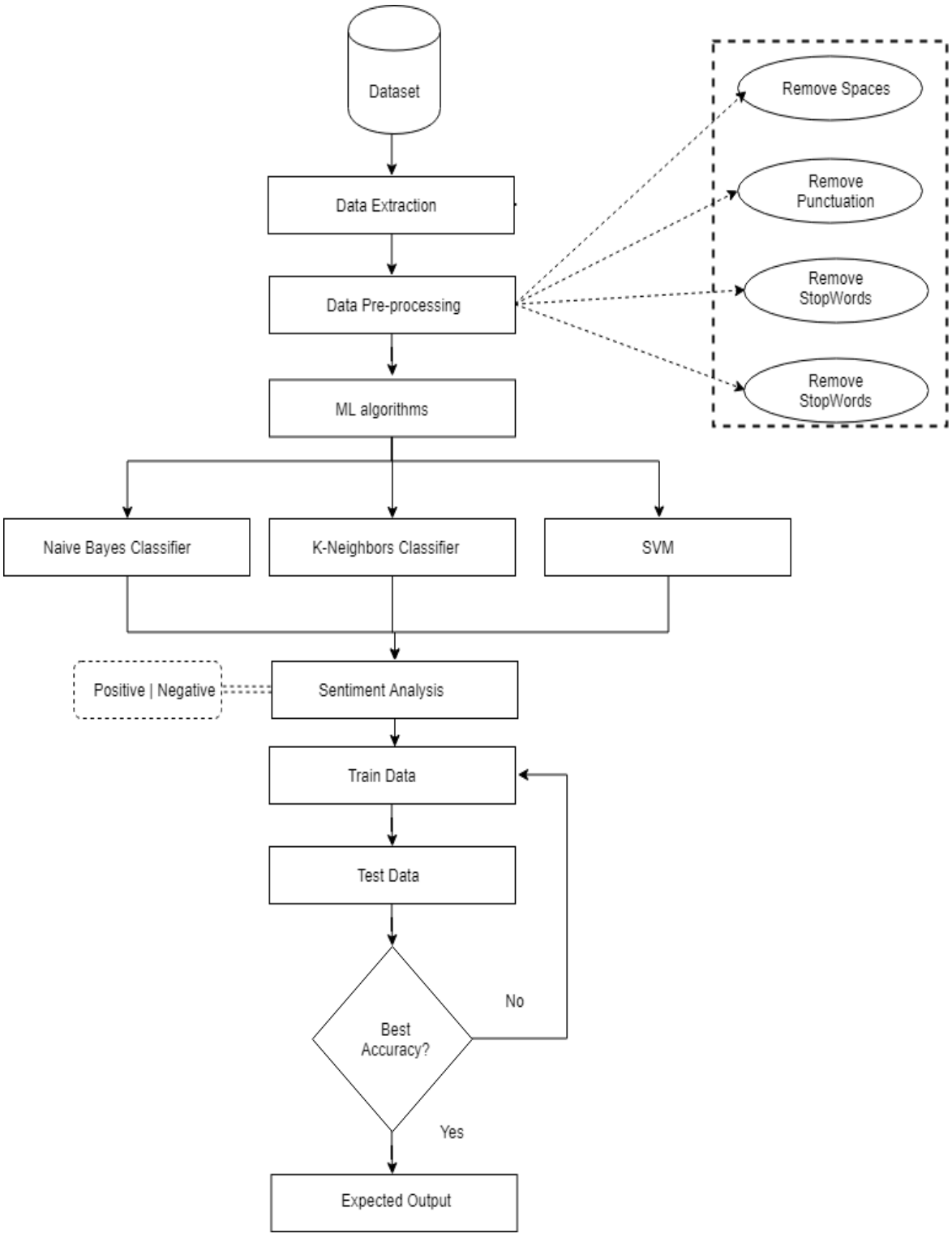


Figure 3.5: Work flow of the model

3.5 Implementation Requirements

It's an important part in this research where we applied some method before data training and testing called data preprocessing.

1) Naïve Bayes Classifier

Naive Bayes classifiers are the collaboration of multiple classifiers algorithm based on Bayes theorem. It is not a single algorithm but it is a collection of sub familiar algorithm where share the common principle. The naïve Bayes algorithm is probabilistic classifiers for classification. It's all feature are independence assumption if compare between predictor. This model is no complicated iterative parameter estimation that's why it's very useful for every big dataset and it is easy to build a model.

Posterior probability P(t|z) is calculating by using Bayes theorem from P(t),P(z) and P(z|t), where P(t) represents the prior probability of class, P(z) is prior probability of predictor and P(z|t) provides the probability the text appears in this class. This forwardness is called class conditional independence.

Consider the following equation is:

$$P(t|z) = \frac{P(t)P(z|t)}{P(z)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where, Class t provides predictor z for posterior probability P(t|z). The value of class t is identified as Positive and Negative where z is a sentence. The value of z is called true when the probability of t is true. because P(t) represents the prior probability of class. P(wi| z) is the probability of the i th feature in given class t where text appears z forget the value of t to maximize P(t|z). We need to train the parameter P(t) and P(wi|t).

2) Support Vector Machine (SVM) Classifier

There are so many different types of the algorithm are available for text classification in machine learning. Support Vector Machines is one of them. We have been chosen this for classification in our experiments. It aligns the text into positive and negative based on the

word. It's can handle the large feature. When the problem is linearly separable and set of example are sparse then it handles them robustly. Vector representation used to encode the information collected from the text which provides SVM. And it also provides a good result classify the text related problem.

SVM is classify following this formula:

$$f(x) = \sum_{i=0}^{n}(a_i k(x, x_i) + \text{b}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where,

$$k(x, x_i) = \exp(-\frac{\|x-x'\|^2}{2\sigma}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

3) K-Nearest Neighbors (K-NN) Classifier

K-nearest neighbor classifier used for classification and regression but there is no need to training for applying this algorithm. It's the commonly used learning algorithm for classification. It's the lazy learning algorithm and non-parametric. So, when a new data set is used for classifying at first it calculates the distance of all point in our data set which is based on K initial value. If K value provides 1 then it merges all of the points which are located nearest distance.

If k>1, It makes a list of K of all data points which represent the minimum distance. It's also made a new data point to follow the maximum data point on the list.

We used a labeled dataset of two peculiarities like we need a new data point for identifying the class which is not labeled data. Then calculate the difference between new data points and all other points. Several distance metrics are available to calculate the distance, it's choosing one of them. Distance function are following:

Eculidean: $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ ..... (4)

Manhattan: $\sum_{i=1}^{n}|x_i - y_i|$ ..... (5)

Minkowski: $(\sum_{i=1}^{n}(|x_i - y_i|)^q)^{1/q}$ ..... (6)

Where, n is no of dimensions, x is data point from our data set and y is predicted new data point.

4) Gradient Boosting Classifier

Gradient Boosting is another process for solving the regression and classification problems which are built a tree one by one frequently and then prediction the output adjustment the summation of an individual tree. It has three component like one is a loss function which is optimized, secondly a weak learner which predict output and then thirdly additive model which add a weak learner to the loss function for minimizing the loss. Here weak learners used for gradient boosting and greedy manner constructed by choosing the best split point. Trees can add one at a single time but the model is not to change. Gradient boosting made the gradient boosting method using various parameter like 'min_samples_leaf','max_depth','random_state','n_estimators',' subsample','learning_rate'. Every testing gets the different combination of values in gradient boost classifier model and then evaluate every combination of accuracy get the best result in this classifier.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

The projected model has taken a special sorts of machine learning algorithmic rule techniques to enhance the accuracy and comparison that one is provides us higher accuracy. This technique tried exhausting for makes higher accuracy. The accuracy of the model was calculated that describe in given below-

4.2 Experimental Results

The setting of the classifier is following:

- **NB:** The batch size is 100 and default probability is 0.
- **KNN:** N-neighbors value is 3, leaf_size is 30 and using metric is 'minknwski'.
- **GB:** N_estimators is 100, learning rate is 1.0,max_depth is 1 and random_state is 0.

For evaluation we used four measured:

- $\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(TN+FN+TP+FP)}$ ..... (7)
- $Precission = \frac{TP}{TP+FP}$ ..... (8)
- $Recall = \frac{TP}{TP+FN}$ ..... (9)
- $F1 - Measure = \frac{2*Precission*Recall}{Precission+Recall}$ ..... (10)

We can see Table 4.1 to get the accuracy and we got the best accuracy is (98.39%) in NB Classifiers comparison all of the others. We also get the highest precision for Negative (N) is (1.0) for KNN and Positive is (0.99) for KNN and NB classifiers. And the best negative recall is (0.11) for NB and Positive recall is (1.0) for SVC, KNN, and GB. Highest Positive F-measure is (1.0) for GB and Negative is (0.19) for KNN. So, our all classifiers accuracy range from 98.26% to 98.39%.

Table 4.1: The result in different classifier

Classifier	TP	FN	FP	TN	Accuracy %	Precision (N,P)	Recall (N,P)	F1-Measure (N,P)
NB	13	104	52	7085	**0.9839**	N - 0.20 P - 0.99	N - 0.11 P - 0.99	N - 0.14 P – 0.99
SVC	13	104	52	7085	**0.9826**	N – 0.0 P – 0.98	N - 0.0 P - 1.0	N - 0.0 P – 0.99
KNN	12	105	0	7137	**0.9835**	N – 1.00 P – 0.99	N – 0.10 P - 1.0	N – 0.19 P – 0.99
GB	0	117	0	7137	**0.9826**	N – 0.0 P – 0.98	N – 0.0 P – 1.0	N – 0.0 P – 1.0

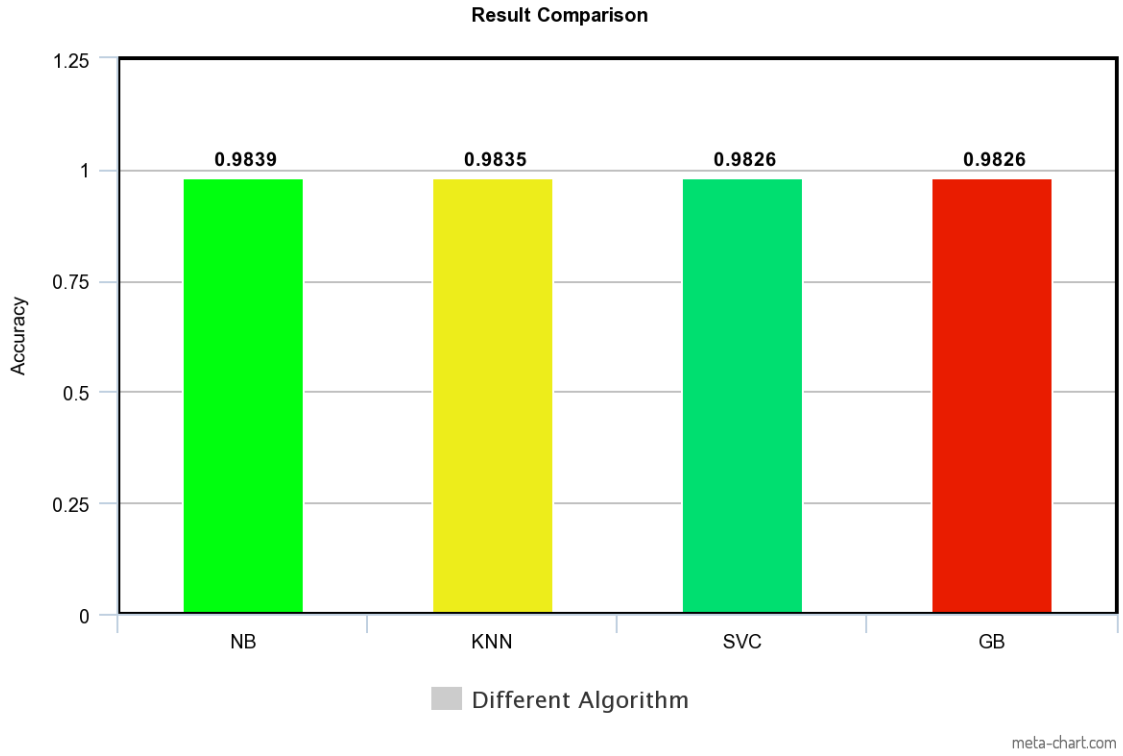


Figure 4.1: Result comparison different algorithm

4.3 Descriptive Analysis

We divided our data set into two section, training dataset included 70% of total data and testing data included 30% of total data. Date set distribution show in Table 1.

Table 4.2: Distribution of dataset

Total Dataset	Training Dataset	Testing Dataset
24178	16924	7254
1.0	70%	30%

4.4 Summary

In this chapter, We addressed our project's accuracy estimation and precision chart. And briefly discussed the result and gave concise explanation.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Summary of the Study

We have implemented a model that is either positive or negative to identify the sentence. We also display the classifier comparison which one is better for an accurate performance. The entire short summer project is given below -

Step 1:

- Data Collection
- Data Pre-processing

Step 2:

- Divided data into train and test set
- Using Different type of Machine Learning algorithm like Naïve Bayes, SVM,K-NN etc.

Step 3:

- Comparison the accuracy which one gives best accuracy.

To analyze the review, these models enhance our business profit. In order to identify human requirements, it also improves social awareness and product quality. The conclusion, recommendations and further improvement ideas of this research will be described in this section.

5.2 Conclusions

Sentiment analysis is a popular topic for data analytics and reporting and advanced processing. In this research main goal is to represent the different type of classifiers and compare their result when classifying types of reviews in real life. The proposed approach must more accuracy if we change the parameter in the classifier. The Naive Bayes classifier feature selection for implementing Emphasizing words handling and negation handling.

5.3 Recommendations

In the next step, we will increase the number of training samples and attempt to use other approaches than comparing neural networks to evaluate what one is doing the best for analyzing sentiment.

It venture has the potential for changes such as:

- This helps people grow their company as an interpreter.
- It is also useful for robotics.
- This can be used by any type of social media channels to exploit user opinion of them.

5.4 Implication for Further Study

We wishes to work a similar type of work in a different language like Bangla especially. Besides, the author wishes more and more experiment execute in Machine Learning algorithm to increase the accuracy. And also wish to include deep learning to analyze the Bengali language sentiment in the future.

References

[1] Zaiuddin, N., & Selamat, A. (2014). Sentiment analysis using Support Vector Machine. *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*. doi: 10.1109/i4ct.2014.6914200

[2] Pang, B., & Lee, L. (2004). A sentimental education. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL 04. doi: 10.3115/1218955.1218990

[3] Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages. *ACM Transactions on Information Systems*, *26*(3), 1–34. doi: 10.1145/1361684.1361685

[4] Zhuang, L., Jing, F., & Zhu, X.-Y. (2006). Movie review mining and summarization. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management - CIKM 06*. doi: 10.1145/1183614.1183625

[5] Chowdhury, S. M. M. H., Abujar, S., Saifuzzaman, M., Ghosh, P., & Hossain, S. A. (2018). Sentiment Prediction Based on Lexical Analysis Using Deep Learning. *Advances in Intelligent Systems and Computing Emerging Technologies in Data Mining and Information Security*, 441–449. doi: 10.1007/978-981-13-1501-5_38

[6] O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C., & Smeaton, A. F. (2009). Topic-dependent sentiment analysis of financial blogs. *Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion - TSA '09*. doi:10.1145/1651461.1651464

[7] Trilla, A., & Alias, F. (2013). Sentence-Based Sentiment Analysis for Expressive Text-to-Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(2), 223–233. doi: 10.1109/tasl.2012.2217129

[8] Ekawati, D., & Khodra, M. L. (2017). Aspect-based sentiment analysis for Indonesian restaurant reviews. *2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*. doi:10.1109/icaicta.2017.8090963

[9] Akhtar, Nadeem, et al. "Aspect Based Sentiment Oriented Summarization of Hotel Reviews." *Procedia Computer Science*, vol. 115, 2017, pp. 563–571., doi:10.1016/j.procs.2017.09.115.

[10] Raghib, Omar, et al. "Emotion Analysis and Speech Signal Processing." *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2017, doi:10.1109/icpcsi.2017.8392246.

[11] Varshney, Vanshika, et al. "Recognising Personality Traits Using Social Media." *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2017, doi:10.1109/icpcsi.2017.8392248.

[12] Bansal, P., Somya, Kamaal, N., Govil, S., Ahmad, T. "Extractive review summarization framework for extracted features" International Journal of Innovative Technology and Exploring Engineering Volume 8, Issue 7C2, May 2019, Pages 434-439

[13] Singh, I., & Sahu, A. K. (2019). A Review on Stone Columns used for Ground Improvement of Soft Soil. *Proceedings of the 4th World Congress on Civil, Structural, and Environmental Engineering*. doi: 10.11159/icgre19.132

b

ORIGINALITY REPORT

5%	3%	3%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.bracu.ac.bd:8080 Internet Source	1%
2	Omar Raghib, Eshita Sharma, Tameem Ahmad, Faisal Alam. "Emotion analysis and speech signal processing", 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017 Publication	1%
3	Doaa Mohey El-Din Mohamed Hussein. "A survey on sentiment analysis challenges", Journal of King Saud University - Engineering Sciences, 2016 Publication	<1%
4	Xujuan Zhou, Yue Xu, Yuefeng Li, Audun Josang, Clive Cox. "The state-of-the-art in personalized recommender systems for social networking", Artificial Intelligence Review, 2011 Publication	<1%
5	"Big Data Analytics and Knowledge Discovery", Springer Science and Business Media LLC,	<1%

	2017 Publication	
6	stuartgeiger.com Internet Source	<1%
7	S. M. Mazharul Hoque Chowdhury, Sheikh Abujar, Mohd. Saifuzzaman, Priyanka Ghosh, Syed Akhter Hossain. "Chapter 38 Sentiment Prediction Based on Lexical Analysis Using Deep Learning", Springer Science and Business Media LLC, 2019 Publication	<1%
8	Arfinda Ilmania, Abdurrahman, Samuel Cahyawijaya, Ayu Purwarianti. "Aspect Detection and Sentiment Classification Using Deep Neural Network for Indonesian Aspect-Based Sentiment Analysis", 2018 International Conference on Asian Language Processing (IALP), 2018 Publication	<1%
9	scholar.uwindsor.ca Internet Source	<1%
10	documents.mx Internet Source	<1%
11	Nadeem Akhtar, Nashez Zubair, Abhishek Kumar, Tameem Ahmad. "Aspect based Sentiment Oriented Summarization of Hotel	<1%

	Reviews", Procedia Computer Science, 2017 Publication	
12	kunz-pc.sce.carleton.ca Internet Source	<1%
13	hal.archives-ouvertes.fr Internet Source	<1%
14	brugtecomputere.dk Internet Source	<1%
15	www.psymbiosys.eu Internet Source	<1%
16	Tata Sutabri, Syopiansyah Jaya Putra, Muhammad Ridwan Effendi, Muhamad Nur Gunawan, Darmawan Napitupulu. "Sentiment Analysis for Popular e-traveling Sites in Indonesia using Naive Bayes", 2018 6th International Conference on Cyber and IT Service Management (CITSM), 2018 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		