An Effective Machine learning Approach for Sentiment Analysis of Restaurant Reviews

Thesis submitted in partial fulfilment of the requirement for the degree of

Bachelor of Computer Science and Engineering

Under the Supervision of

Dr. Jia Uddin

By

Rabita Karim (13101248)

School of Engineering and Computer Science

August 2016

BRAC University, Dhaka, Bangladesh

Declaration

I hereby declare that this thesis is based on results obtained from our own work. Due acknowledgement has been made in the text to all other material used. This thesis, neither in whole nor in part, has been previously submitted to any other University or Institute for the award of any degree or diploma.

Signature of Supervisor **Signature of the Authors**

Dr. Jia Uddin Rabita Karim(13101248)

Acknowledgement

First and foremost, I would like to thank Almighty Allah for enabling us to initiate the research, to put our best efforts and successfully conclude it.

Secondly, we submit our heartiest gratitude to our respected Supervisor Dr. Jia Uddin for his contribution, guidance and support in conducting the research and preparation of the report. Every last involvement of his, starting from instilling in us the deadliest of fears to the kindest words of inspiration has permitted us to effectively complete the paper. I am truly grateful to him.

We revere the patronage and moral support extended with love, by our parents as well as our friends. They helped us with their direct or indirect suggestions which aided in achieving our goal. We would also like to acknowledge the assistance we received from numerous resources over the Internet especially from fellow researchers' work.

Last but not the least, I thank BRAC University for providing us the opportunity of conducting this research and for giving us the chance to complete our Bachelor degree.

Table of Contents

List of Figures

List of Tables

List of Acronyms

A. ML–Machine Learning

B. GNB–Gaussian Naive Bayes

C. DT–Decision Tree

Abstract

Sentiment analysis or opinion analysis is creating a vast area of research in this modern era of social media. Various blogs and Social Medias (Facebook, twitter, Instagram) are the most popular platform for the users or consumers where they frequently express their opinion about current topics, various brands, restaurants, movies, books, traveling places etc. Sentiment analysis is a very smart and effective approach to find peoples view about a particular news/ place/restaurant/movie/book/brand. It is beneficial for the both service providers or sellers and consumers. Researchers in the areas of natural language processing, data mining, machine learning, and others have tested a variety of methods of automating the sentiment analysis process. In this research work, I used restaurant reviews dataset to analysis the sentiment and for this approach Gaussian Naïve Bayes method is proposed based on coupling classification methods using arcing classifier and their performances are analyzed in terms of accuracy.

Chapter 1

1. Introduction

1.1 Introduction

In this era of globalization we are always looking for some idea that can save our time, reduce the task complexity, replacement of manual processing. Sentiment analysis is such a novel approach by which we can easily save our time, make our task easy. Sentiment analysis is a fancy word that data scientists use instead of emotion detection. Sentiment or opinion mining refers to the type of natural language processing used to understand the moods, opinions and sentiments of the public regarding a particular product or a movie or an event. Machine learning is the one of the most prevalent approach to analysis sentiment. In this paper we used Gaussian Naïve Bayes algorithm and Decision Tree algorithm as machine learning classifier.

According to the Yelp (an American multinational corporation) Dataset Challenge reviews dataset contains 1,569,264 business reviews. The most prominent category for reviews is Restaurants with 990,627 restaurant reviews. That's why, I used restaurant review for sentiment analysis [3].

1.2 Brief about Sentiment Analysis

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. This writing could be a tweet, book review, movie review, restaurant review, brand review, place review, product review etc. It can be used to identify the customer or follower's attitude towards a brand through the use of variables such as context, tone, emotion etc.

These types of reviews are equally important and helpful for the both consumers and service providers. As a consumer, other consumers review is helpful for him to get the idea of a product. On the other hand, service providers can check consumers' review about their product to determine the acceptance of their products. Marketers can use sentiment analysis to research public opinion of their company and products, or to analyze customer satisfaction. Organizations can also use this analysis to gather critical feedback about problems in newly released products.

It is a very time consuming approach to go through a huge numbers of reviews manually about any product or service. Sentiment analysis is an effective replacement of this manual review reading process. By using sentiment analysis approach it can be easily detect within a short amount of time, whether a product is liked by customers or not.

Sentiment analysis not only helps companies understand how they're doing with their customers, it also gives them a better picture of how they stack up against their competitors. For example, if 'X'company has 20% negative sentiment, is that bad? It depends. If its competitors have a roughly 50% positive and 10% negative sentiment, while X's is 20% negative, that merits more discovery to understand the drivers of these opinions. Knowing the sentiments associated with competitors helps companies evaluate their own performance and search for ways to improve [1].

1.3 Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves [2].

There are various machine learning algorithm. Among them we worked with Gaussian Naïve Bayes (GNB) and Decision Tree (DT) algorithm. We basically use Gaussian Naïve Bayes (GNB) classifier and Decision Tree (DT) classifier to fit it to the training dataset. We use python programming language to build our model. As an editor, we use Anaconda 4.2.

1.4 Motivation and Background Studies

Sentiment analysis, also called opinion mining, is a form of information extraction from text of growing research and commercial interest. A number of researchers tried to research on this topic and there are already large numbers of papers published in this topic. Our work is actually motivated from the recent advancements in this machine learning techniques. We can achieve sentiment analysis in different levels as word, sentence and document levels. In many of the reviews people express diverse opinion about features. So, aspect/feature based sentiment analysis would be the most suitable for our work. Apart from this we work on finding the sentiment of restaurant based data which will help the restaurant owners.

There are large numbers of works on related topics such as G and et al., (2005) presents us an overview of recommend systems. In this work they describe the current version of recommendation methods which are mainly divided into three categories, content-based, collaborative and hybrid recommendation approaches. However, there are limitations on these approaches [12]. This paper discusses several possible extensions that can improve recommendation capabilities, as well as make recommendation systems applicable to a broader range of application.

Twitter data has been vastly used the era of sentiment analysis research. A numbers of researchers work with twitter dataset as twitter is one of the biggest microblogging service on the internet. Bac Le and et al, (2015) used 200000 tweets and proposes a sentiment analysis model based on Naive Bayes and Support Vector Machine [4].By using standard machine learning techniques Ziqiong Zhang and et al., (2011) and using the naive Bayes and SVM algorithms which are incorporated into the domain of online Cantonese-written restaurant reviews to automatically classify user reviews as positive or negative. The effects of feature presentations and feature sizes on classification performance are discussed in the work. Gayatree Ganu and et al., (2009) gave us a more similar example in their work. A free-text format review is difficult for computers to analyze, understand and aggregate. To identify the information in the text reviews, this paper propose a new ad-hoc and regression-based recommendation methods that takes into consideration the textual component of user reviews [5].

Chapter 2

Proposed Model

In this chapter, a system model of Machine learning is presented. The model consists of seven major components along with few more sub components.

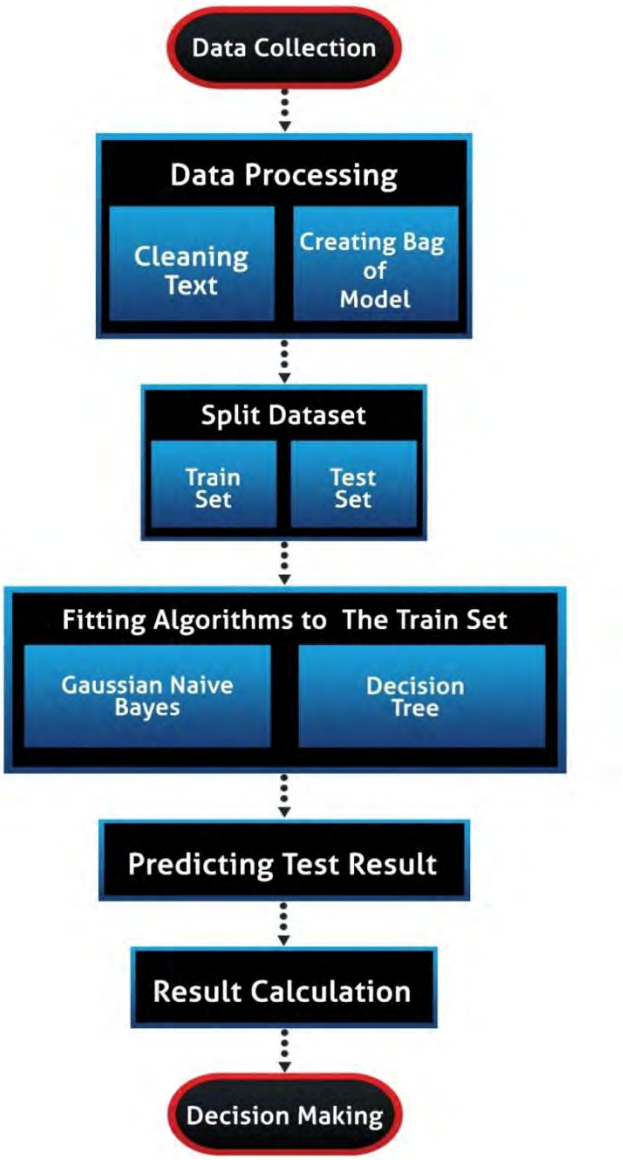

Figure 01: Proposed Model

2.1 Data collection

In this work, we used restaurant reviews as our reference dataset. From www.superdatascience.com website, we collected the data where 1000 reviews of a restaurant and its corresponding sentiments are provided. This dataset has two columns. First column contains the text data and second column contains the sentiment which is representing by binary values. Such as if a review is in the favor of restaurant as in if it is a good review then, its corresponding sentiment is defined as "1". On the other hand, if a review is not in the favor of restaurants as in bad review is defined as "0".

Table 01: Glimpse of Dataset

Review	Liked
Wow... Loved this place.	1
Crust is not good.	0
Not tasty and the texture was just nasty.	0
Stopped by during the late May bank holiday off Rick Steve recommendation and loved it.	1
The selection on the menu was great and so were the prices.	1
Now I am getting angry and I want my damn pho.	0
Honeslty it didn't taste THAT fresh.)	0
The potatoes were like rubber and you could tell they had been made up ahead of time being kept under a warmer.	0
The fries were great too.	1
A great touch.	1
Service was very prompt.	1
Would not go back.	0
The cashier had no care what so ever on what I had to say it still ended up being wayyy overpriced.	0
I tried the Cape Cod ravoli, chicken, with cranberry...mmmm!	1
I was disgusted because I was pretty sure that was human hair.	0
I was shocked because no signs indicate cash only.	0
Highly recommended.	1
Waitress was a little slow in service.	0
This place is not worth your time, let alone Vegas.	0
did not like at all.	0
The Burrittos Blah!	0

2.2 Data preprocessing

As we are dealing with text data for sentiment analysis, data preprocessing plays a vital role on our research to make the model understand the data. Text data contains a lot of noise. As a result it's a challenge to clean the texts smartly

Data pre-processing reduces the size of the input text documents significantly. It happens by various steps.

Import Regular Expression: Regular expression is sequence of character mainly ised to find and replace patterns in a string,

Stop-word elimination: Stop-words are functional words which occur frequently in the language of the text (for example, „a", "the", "an", "of" etc. in English language), so that they are not useful for classification. We use here Natural Language Toolkit (nltk) library for the reduce stop-words.

Stemming: In information retrieval, stemming is the process of reducing inflected words to their root (or stem), so that related words map to the same stem. This process naturally reduces the number of words associated with each document, thus simplifying the feature space. In our experiments we use an implementation of the Porter stemming algorithm [7]. For example, using the Porter"s stemmer, the English word "generalizations" would subsequently be stemmed as "generalizations → generalization → generalize → general → gener".

Bag of model: Bag of words Model, basically in the first big step of natural language processing we not only cleaned all the review, but also created a corpus. Corpus refers the collections of texts. In our model our corresponding corpus contains 1000 of cleaned reviews.

From the corpus, we make the bag of model. Bag of model takes all the different words from the corpus. In our corpus, there is 1000 reviews and for every different words from these reviews, it takes one corresponding column. 1000 reviews contain a lot of different words, that's why it contains a lot of columns. We put all columns in a table and row number is 1000.

By using Bag of Model, we can easily avoid duplicate values, data redundancy. Each sell contains number and this number is going to be the number of times the corresponding to the column appears in the review. For example, first column is “wow love place”, so we’ll get 1 for ‘wow’ cell, in the 2nd row, there is ‘wow’ is not present, so for the 2nd row, the value of wow cell will be ‘0’. This is the approach of creating the Bag of Model.

Table 02: A portion of bag of model

Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

2.3 Split Dataset

One of the important parts of Machine Learning model is to split the data set into two parts. These are-

(a) Train set

(b) Test Set

The fundamental goal of Machine Learning is to *generalize* beyond the data instances used to train models. We want to test the model to estimate the quality of its pattern generalization for data the model has not been trained on. However, because future instances have unknown target values and we cannot check the accuracy of our predictions for future instances now, we need to

use some of the data that we already know the answer for as a proxy for future data that is called our Test Set.

In case large data, the most usual way of split it into training and test subsets, usually with a ratio of 70-80 percent for training and 20-30 percent for test set. This process of splitting is done randomly by Train_test_split function which is imported from scikit-learn libraryfunction.

Train Set

We have put 80% data as in 1600 reviews into our train set. In the training set, both the independent variables (X_train matrix) and dependent variables (Y_train matrix) are known

Test Set:

Left 20% data as in 40 reviews are in test, where dependent variables are known as X_test matrix and independent variables are known Y_test matrix.

Cross Validation techniques:

Cross-validation techniques belong to conventional approaches used to ensure good generalization and to avoid over-training. We use cross validation fromscikit-learn library. Cross-validation techniques can also be usedwhen evaluating and mutually comparing more models, various training algorithms, or when seeking for optimal model parameters. Our main purpose of using cross-validation is to achievea stable and confident estimate of the model performance.

Chapter 3

Fitting Algorithm to Train Set

It is difficult to find a good or even a well-performing machine learning algorithm for a particular dataset.

We went through a process of trial and error to settle on a short list of algorithms that provides better result. We studied a couple of algorithms. In our work we are going to show and discuss the performance of Gaussian Naive Bayes and Decision Tree algorithms.

3.1 Predict the Test Result

The Machine Learning system uses the training data to train models to see patterns, and uses the test data to evaluate the predictive quality of the trained model. The ML system evaluates predictive performance by comparing predictions on the evaluation data set with true values (known as ground truth) using a variety of metrics.

By Gaussian Naive Bayes and Decision Tree algorithms, we are going to predict the test result as in the value of Y_pred.

3.2 Machine Learning Classification Model (Classifier)

In the terminology of machine learning, classification is considered an instance of supervised learning, learning where a training set of correctly identified observations is available. [6]

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm that maps input data to a category.

We use Gaussian Naive Bayes classifier and Decision Tree classifier to predict the restaurant reviews whether it is good or bad.

3.3 Gaussian Naive Bayes classifier

In machine learning, naive Bayes classifier is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

3.3.1 Why Gaussian Naive Bayes

There are some instinct points behind choosing Gaussian Naive Bayes

First of all, Naive Bayes is a classification algorithm suitable for binary and multiclass classification. In our dataset, the sentiments are representing as binary value.

If the input variables are real-valued, a Gaussian distribution is assumed. In which case the algorithm will perform better. Our data set is a real valued dataset.

Text classification/ Spam Filtering/ Sentiment Analysis:

Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments). In this thesis, our aim is analysis the sentiment by using the dataset of restaurant reviews.

3.3.2 Gaussian Naive Bayes

This extension of naive Bayes is called Gaussian Naive Bayes which is used for real- valued attributes. Other functions can be used to estimate the distribution of the data, but the Gaussian (or Normal distribution) is the easiest to work in case where only the mean and the standard deviation from training data is need to be estimated.

3.3.3 Representation for Gaussian Naive Bayes

With real-valued inputs, Gaussian Naive Bayes can calculate the mean and standard deviation of input values (x) for each class to summarize the distribution.

This means that in addition to the probabilities for each class, we must also store the mean and standard deviations for each input variable for each class.

3.3.4 Learn a Gaussian Naive Bayes Model from Data

This is as simple as calculating the mean and standard deviation values of each input variable (x) for each class value.

$$Mean(\mu_y) = \frac{\sum x_i}{n}$$

Where n is the number of instances and x are the values for an input variable in atraining data.

We can calculate the standard deviation using the following equation:

Standard Deviation (σ)= $\sqrt{\frac{1}{n} \times \sum(x_i - \mu_y^2)}$

This is the square root of the average squared difference of each value of x from the mean value of x, where n is the number of instances function, xi is a specific value of the x variable for the i'th instance and $\mu_y$is the mean.

3.3.5 Make Predictions with a Gaussian Naive Bayes Model

Probabilities of new x values are calculated using the Gaussian Probability Density Function$P(x_i|y)$

When making predictions these parameters can be plugged into the Gaussian PDF with a new input for the variable, and in return the Gaussian PDF will provide an estimate of the probability of that new input value for that class. [8]

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2})$$

Where,$P(x_i|y)$ is a Gaussian Probability Density Function.

3.3.6 Gaussian Naive Bayes Classifier

We need to import pandas, numpy and sklearn libraries. From sklearn, we need to import preprocessing modules like Imputer.

We have built a **GaussianNB** classifier. The classifier is trained using training data. We can use **fit ()** method for training it. After building a classifier, our model is ready to make predictions. We can use **predict ()** method with test set features as its parameters.

3.4 Decision Tree Algorithm

A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node.

Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items.

Decision Tree Algorithm Pseudo code

Place the best attribute of the dataset at the root of the tree.

Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.

Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

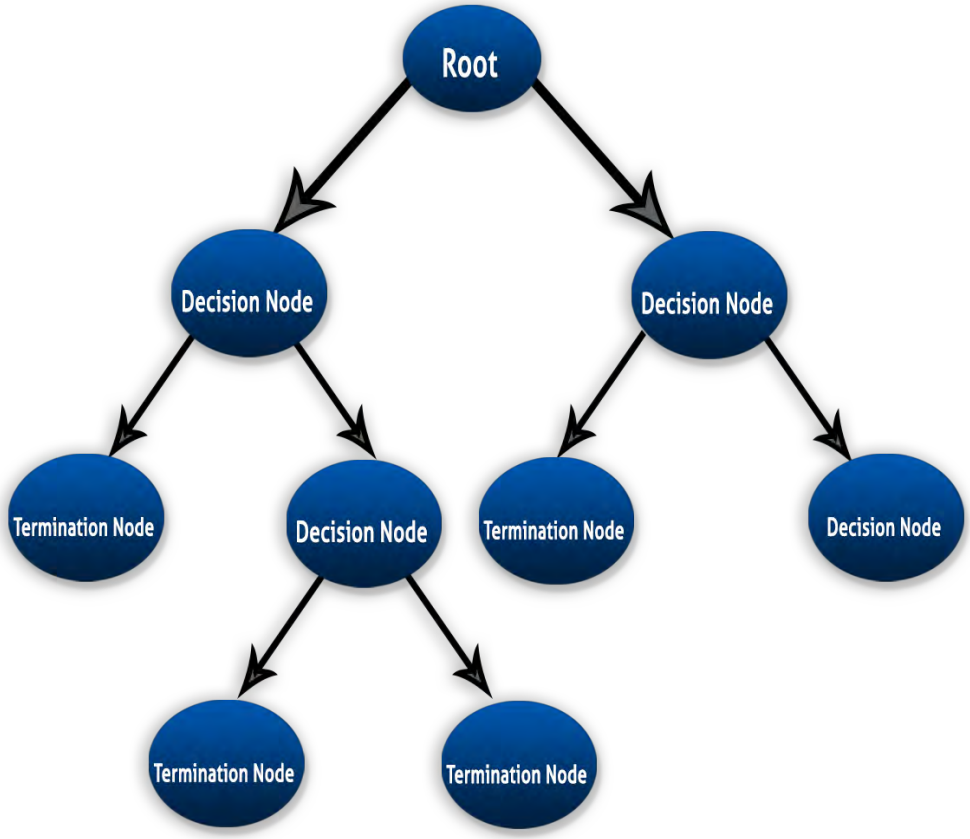

Figure 02: Decision Tree

In decision trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

We continue comparing our record's attribute values with other internal nodes of the tree until we reach a leaf node with predicted class value. As we know how the modeled decision tree can be used to predict the target class or the value. Now let's understanding how we can create the decision tree model.

3.4.1 Decision Tree Classifier

(a) Import Library

(b) Import other necessary libraries like pandas, numpy...

(c) from sklearn import tree

(d) Assumed you have, X (predictor) and Y (target) for training data set and _test(predictor) of test_dataset

(e) Create tree object

(f) model = tree.DecisionTreeClassifier(criterion='gini') # for classification, here you can change the algorithm as gini or entropy (information gain) by default it is gini

(g) model = tree.DecisionTreeRegressor() for regression

(h) Train the model using the training sets and check score

(i) model.fit(X, y)

(j) model.score(X, y)

(k) Predict Output

(l) predicted= model.predict(x_test)

Chapter 4

Result Calculation

Two classification methods (Gaussian Naïve Bayes and Decision Tree) are adapted for each training set. The results are evaluated using the cross validation method on restaurant review based on the classification accuracy.

4.1 Result of Gaussian Naive Bayes (GNB) Model

After applying Gaussian Naive Bayes (GNB) classification Model, we get the sentiment of those reviews which are structured in the X_test matrix.

The below table represents some of our actual sentiment of test set and predicted sentiment that we get aft

Table 03: Actual Sentiment and GNB predicted Sentiment

Actual Sentiment (Y_test)	Predicted Sentiment by GNB (Y_Pred)
0.00	1.00
0.00	1.00
0.00	1.00
0.00	0.00
0.00	0.00
0.00	1.00
1.00	1.00
0.00	1.00
0.00	1.00
1.00	1.00
1.00	1.00
1.00	1.00
0.00	1.00
1.00	1.00
1.00	1.00
1.00	1.00
0.00	0.00
0.00	0.00
0.00	0.00

This table represents some of our actual sentiment of test set and predicted sentiment that we get after the application DT classification.

Table 04: Comparison of Actual Sentiment and GNB predicted Sentiment

Sentiment	Bad Review(0)	Good Review(1)
Actual Sentiment(Y_test)	97	103
Predicted Sentiment by GNB (Y_pred)	67	133

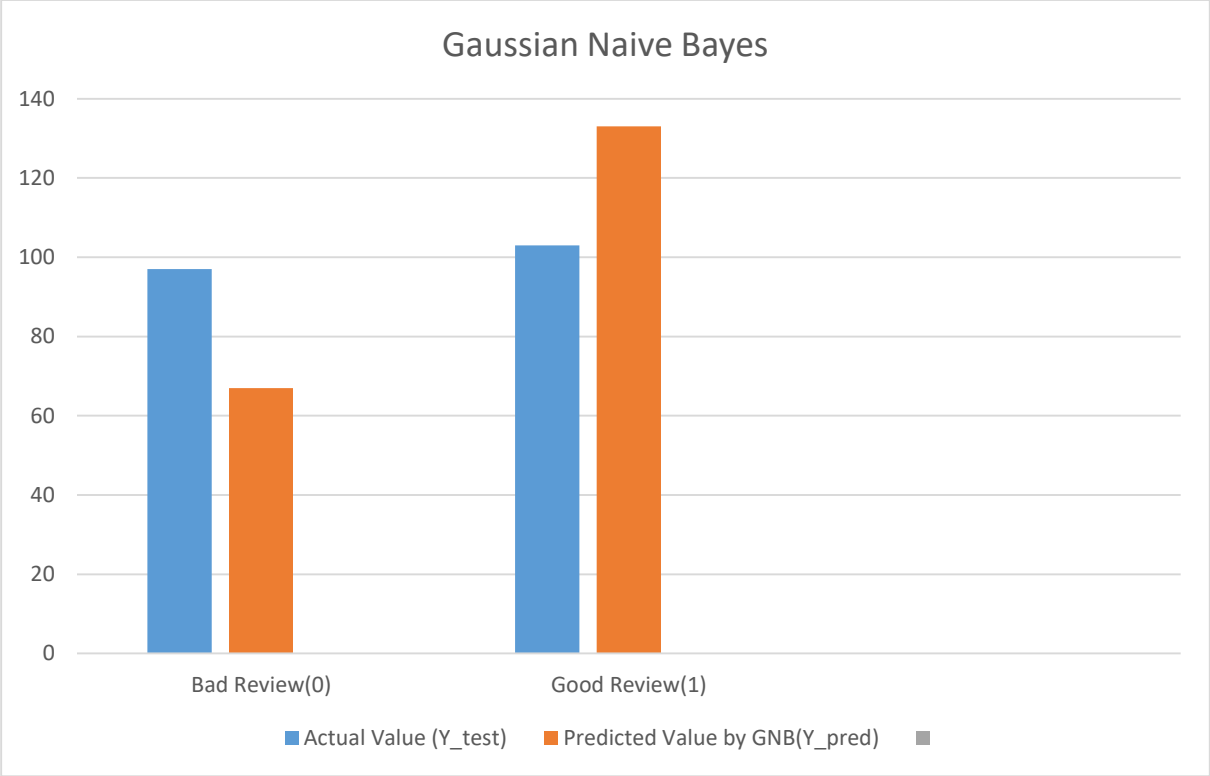

Figure 03: Graph of Actual Sentiment and GNB predicted Sentiment

After fitting GNB model to the test set, we find the Classification Accuracy: 0.7299 ≈ 73%

4.2 Result of Decision Tree Model

This table represents some of our actual sentiment of test set and predicted sentiment that we get after the application DT classification.

Table 05: Actual Sentiment and DT predicted Sentiment

Actual Sentiment (Y_Test)	Predicted Sentiment (Y_Pred)
0.00	0.00
0.00	0.00
0.00	1.00
0.00	0.00
0.00	1.00
0.00	1.00
1.00	1.00
0.00	1.00
0.00	0.00
1.00	1.00
1.00	1.00
1.00	1.00
0.00	1.00
1.00	1.00
1.00	1.00
1.00	1.00
0.00	0.00
0.00	0.00
0.00	0.00
1.00	0.00
0.00	0.00

Table 06: Comparison of Actual Sentiment and DT predicted Sentiment

	Bad Review(0)	**Good Reviw(1)**
Actual Value (Y_test)	**97**	**103**
Predicted Value by GNB (Y_pred)	**67**	**133**

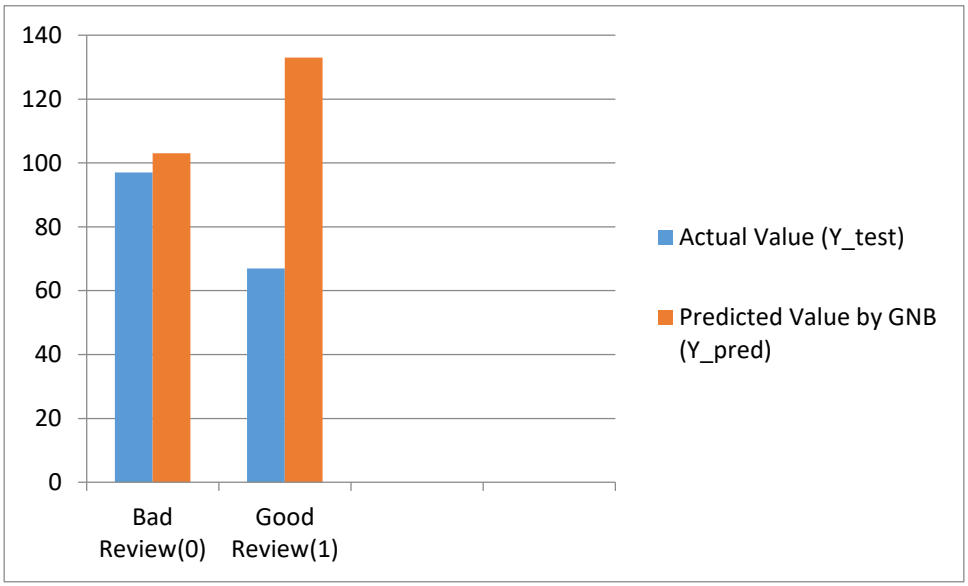

Figure 04: Graph of Actual Sentiment and DT predicted Sentiment

Classification Accuracy: 0.64500000000000002≈ 65%

4.3 Performance Analysis and Decision Making

From our Result Calculation chapter, we get accuracy around 73% from our Gaussian Naïve Bayes Classifier Model and around 65% accuracy from our Decision Tree Classifier Model.

In this chapter, we will discuss both this result and will come to a decision to propose a better model for sentiment analysis of our given restaurant reviews.

We will elaborately discuss the total performance of both GNB classifier and DT Classifier to come to a decision.

Total Performance of GNB Classifier:

Table 07: Total performance of GNB Model

	Bad(0)	Good(1)
Actual	97	103
Predicted Sentiment	67	133
Right Sentiment	55	91
Wrong Sentiment	8	42

From this table, we can see that, among 97 bad reviews GNB perfectly predicted 55 reviews.

On the hand among 103 good reviews, GNB perfectly predicted91 reviews. The average percentage of its prediction is 73% which is near to our classification accuracy of GNB model 0.7299.

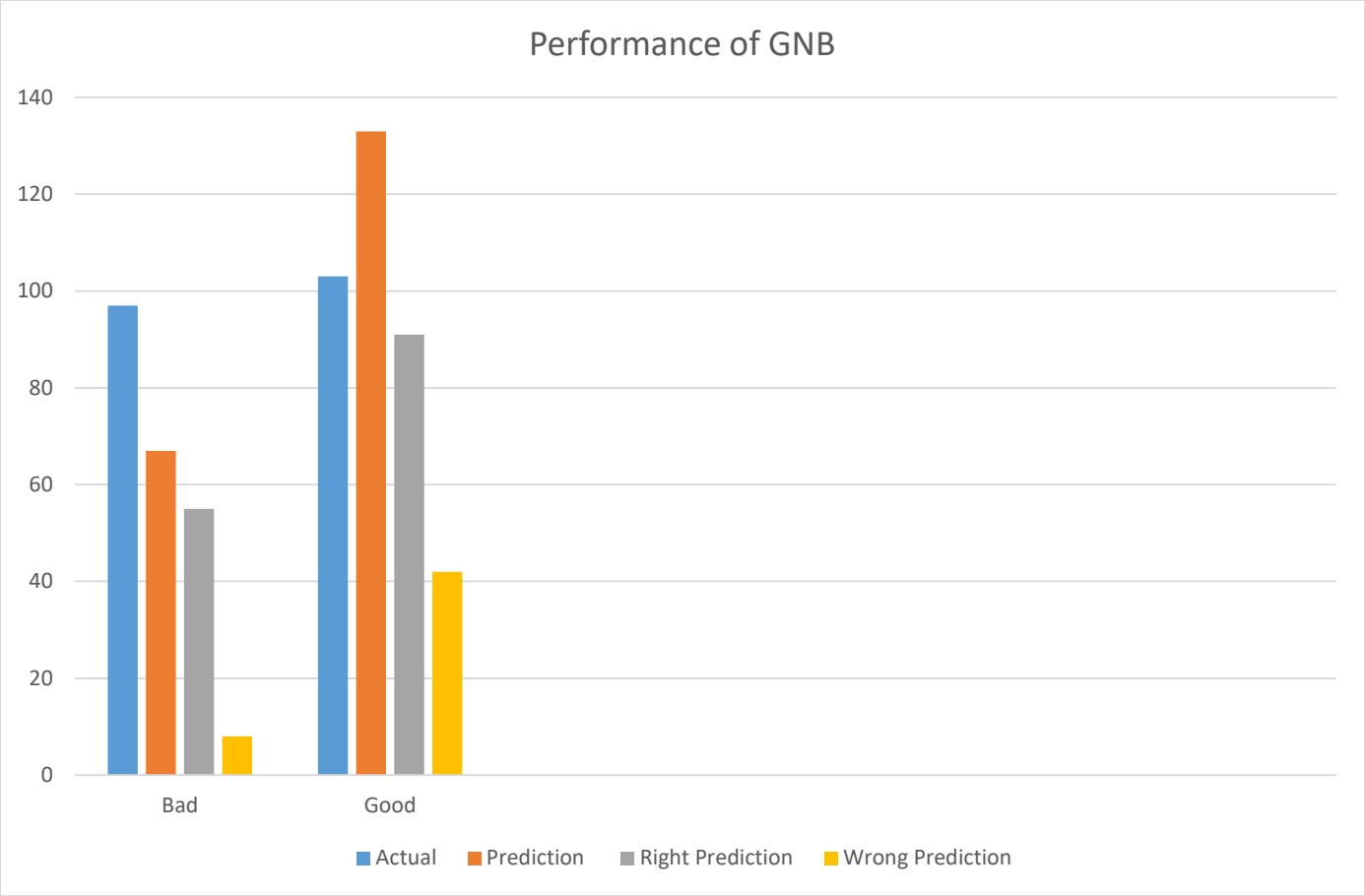

Figure 05: Total Graph representation of GNB predicted Sentiment

Total Performance of DT Classifier:

Table 08: Total performance of DT Model

	Bad(0)	Good(1)
Actual	97	103
Predicted Sentiment	108	92
Right Sentiment	61	69
Wrong Sentiment	47	37

From this table, we can see that, among 97 bad reviews DT perfectly predicted 61 reviews. On the hand among 103 good reviews, DT perfectly predicted 69 reviews. The average percentage of its prediction is 65% which is near to our classification accuracy of DT model 0.6450.

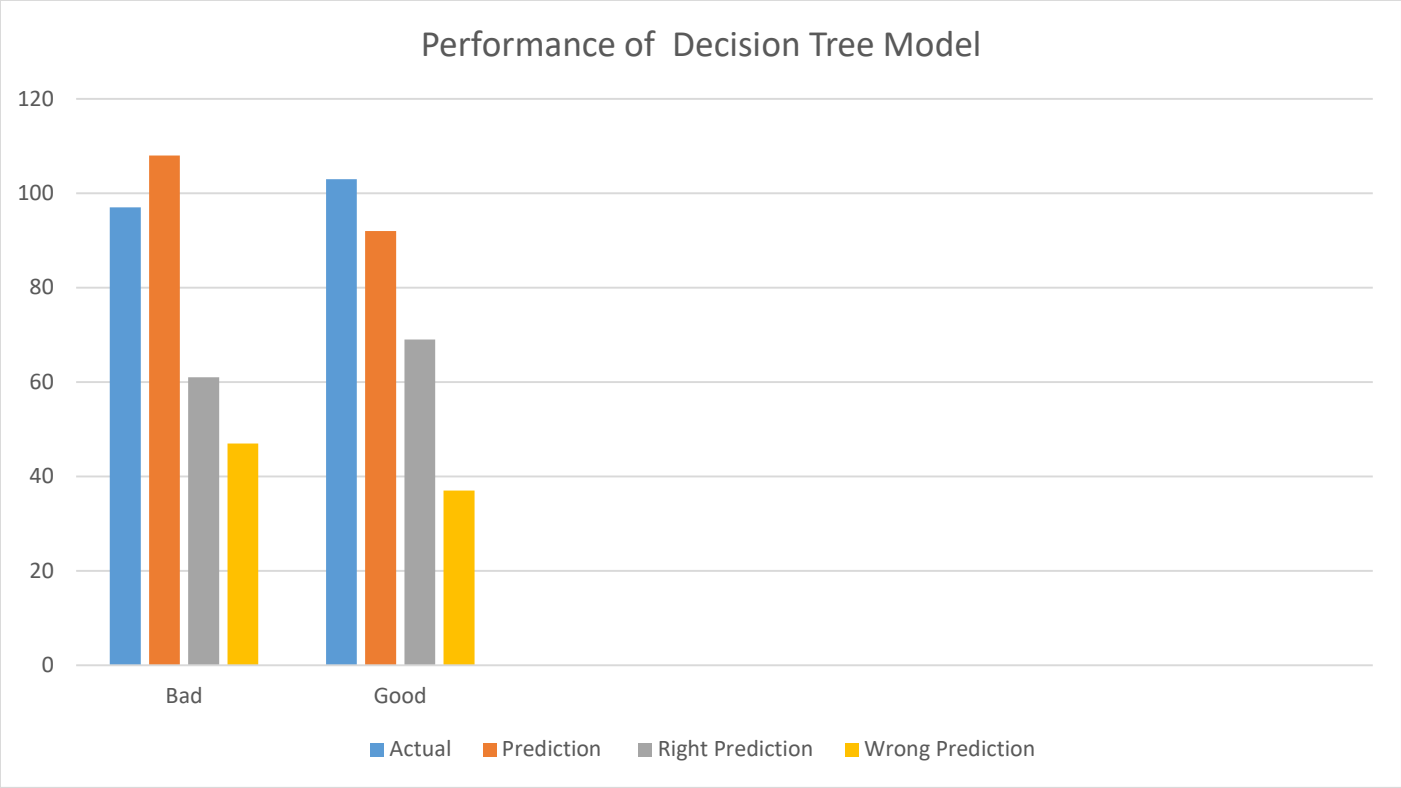

Figure 06: Total Graph representation of GNB predicted Sentiment

From this we can come to decision that, Gaussian Naïve Bayes Classifier is giving the best performance for our model of sentiment analysis. It is predicting 65% sentiment correctly from the restaurant review dataset.

Chapter 5

Conclusion

In this thesis, we have proposed Gaussian Naïve Byes Classifier Model for sentiment analysis

This model can be used for sentiment analysis any kind of text data such as tweets, brand/ product review, travel place review. We have apply this model on a dataset that contains 2000 restaurant reviews.

Sentiment analysis is very much important for both consumers and service providers. Now this modern era of web and globalization, both consumers and service providers are interested to know the overall opinion of people about a particular bran/product/ place etc. It is beneficial for the service provider as it has some business aspect, at the same time it is beneficial for the consumers as it helps them choose the best product.

From our thesis work we have come to a decision that Gaussian Classifier is an effective machine learning model for sentiment analysis. It provides a better prediction for sentiment analysis.

In the field of sentiment analysis it is a big challenge to analysis the sarcastic review/text. Machine can to detect sarcasm. Future research can focus on sarcastic expressions which are usually difficult to understand, both by the users' and the computer system. One more challenging issue is the detection of spam contents in users' review. Finally the study can be extended to resolve the problem of co-reference resolution.

Reference

[1] IPULLRANK blog, http://ipullrank.com/step-step-twitter-sentiment-analysis-visualizing-united-airlines-pr-crisis/, accessed on 10th August, 2017.

[2] Expert System Semantic Intelligence,http://www.expertsystem.com/machine-learning-definition/, accessed on 10th August, 2017.

[3] RPubs brought to you by RStudio, http://rpubs.com/yceron/155272, accessed on 9th August, 2017.

[4] B. Le, H. Nguyen, "Twitter Sentiment Analysis Using Machine Learning Techniques", "Springer", 2015.

[5] G. Ganu, N. Elhadad, and A. Marian, "Association for Computational Linguistics",beyond the stars: Improving rating predictions using review text content. In WebDB, volume 9, pages1–6.2009.

[6] E. Alpaydin"Introduction to Machine Learning", "MIT Press", 2010.

[7] M. F. Porter, "Readings in Information Retrieval," K. Sparck Jones and P. Willett, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, ch. An algorithm for suffix stripping, pp. 313–316.

[8] J. Brownlee, "Naive Bayes for Machine Learning", "Machine Learning Mastery", 2016.

[9] M. Govindarajan, "Sentiment Analysis of Restaurant Reviews Using Hybrid Classification Model", "IRF International Conference", 9th February 2014, Chennai India.

[10] B. Kharadi, K. Patel ,"Opinion Mining of Restaurant Review by sentiment Analysis Using SVM", "International Journal of Innovative Research in Computer and Communication Engineering", 2017, India.

[11] M. Ayyavaraiah, "Review of Machine Learning based Sentiment Analysis on Social Web Data", "International Journal of Innovative Research in Computer and Communication Engineering", 2016, India.

[12] E. Boiy, M. F. Moens," A machine learning approach to sentiment analysis in multilingual Web texts", Springer- Information Retrieval Journal, 2008, Belgium.

[13] D. V. N. Devi, C. K. Kumar, S. Prasasd ,"A feature Based Approach for Sentiment Analysis by Using Support Vector Machine", "IEEE 6th International Conference on Advanced Computing", 2016, India .