"MULTICLASS PREDICTION BASED ON TRANSPORT REVIEW USING DATA MINING ALOGORITHM"

BY

MD SADIKUL ISLAM
ID: 133-15-2924

AND

MD MAHAMUD HASAN
ID: 142-15-4064

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

Supervised By

Ms. Nasrin Akter
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
MAY 2018

APPROVAL

This Project titled **"Multiclass Prediction Based on Transport Review using Data Mining Algorithm**", Submitted by **Md. Sadikul Islam** and **Md. Mahamud Hasan** 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 (BSc) and approved as to its style and contents. The presentation has been held on 08.04.2018.

BOARD OF EXAMINERS

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

Dr. Sheak Rashed Haider Noori — **Internal Examiner**
Associate Professor and Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Md. Zahid Hasan — **Internal Examiner**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

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

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Ms. Nasrin Akter, 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. NASRIN AKTER
Lecturer
Department of CSE
Daffodil International University

CO-SUPERVISED BY:

MR. SAIFUL ISLAM
Lecturer
Department of CSE
Daffodil International University

SUBMITTED BY:

(MD SADIKUL ISLAM)
ID: 133-15-2924
Department of CSE
Daffodil International University

(MD MAHAMUD HASAN)
ID: 142-15-4064
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 Thesis successfully.

We really grateful and wish our profound our indebtedness to **Supervisor Ms. Nasrin Akter, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining*" 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 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

In recent years it is noticeable that sharing text reviews on various businesses specially restaurants through website and social media is a very common phenomenon. Online reviews reflect user's opinion. This huge collection of user data in terms of text reviews can be analyzed to identify user's sentiment and their demand also. Here users are the primary sources. Text reviews are the complete reflection of user's sentiment and also owned by them. Measuring user's sentiment will also be able to find out the market position of a Transportation system. By making the machine learned about the total reviews, it will be able to categorize the unknown text. We collect the necessary data for our research work from a verified source and Google Play store. We took a step forward by combining user review texts which were collected from that website to build a model that can predict a review asserting good or bad and Average. Key benefit of our approach is that, by using our proposed model transport system Owners can identify the main focused term from the review of customers and also can take future step to work on that.

We are also able to publish the position of a System by counting that how many reviews are Good , Bad, and Average comparative to with each. As this model is based on text document, it will be very perfect work in all terms and condition. Because text document shows almost the best predicting result of user's sentiment than that of star rating does.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES:

LIST OF TABLES

TABLES

LIST OF ABBREVIATION

DIU – Daffodil International University

CSE – Computer Science and Engineering

ML – Machine Learning

RQ – Research Question

NLP – Natural Language Processing

POS – Parts Of Speech

NLTK – Natural Language Toolkit

GPL – Google Play Store

API – Application Programming Interface

SVM – Support Vector Machine

KNN – K Nearest Neighbors

CHAPTER 1
INTRODUCTION

1.1 Introduction

In the latest years it's promised that sharing text review on various platform on internet like different website and the social networking sites is very common affair. Now a day's people started to share their feeling by posting in different social site even different types of transport related site and the related social media pages in the internet. Online statement presents the actual mind state of a person. By providing several types statement users actually expresses their opinion about their favorite transport system. With the passing of the time Everything's in the word are getting changed. In this modern time the communication system also changing. In this time people can go very easily here and there through the modern Transportation way.

In respect of technological and modern evolution 'Pathao' and 'Uber', the Bangladeshi new transport system with mobile application is giving a new dimension of the public Transportation. Actually, the whole system is being controlled by a mobile application. It is remarkable that in the current time, the number of internet user are increasing drastically. Maximum of the percentage of people are familiar with the internet browsing and they use it in various purposes. At times they used it to considering different types of review in many sites as they concern. Actually, the online review's work like the mirror of mind of a person. There has a big possibility to identify their sentiment by analysis these reviews what they are considering. In this circumstance, these all reviews can be a great source of user's sentiment. The Bangladeshi Local transportation system the "Pathao" and the "Uber" owner can offer their users to share their experience by using review in social media and their website so that the owner begin to care about their user's key factor of interest approximately their services.

In this research work, We took a step forward through combining users texts review from 'Pathao' and "Uber" application in google play store which is totally own by Bangladesh By doing the machine learned with the training dataset then it is possible to categorize totally unknown text [2].

As our model is based on a textual content record, so it will be prefect work in all terms and condition. Because text report indicates nearly the best predicting result of persons Sentiment. We

select all the valid text which only are in English. Then we build a supervised Learning model to predict the level of assessment text.

In this study, we took a step by way of combining person evaluate text review which is collected from the google play store. By building a model that may expect an overview of Good, Bad and Average [1,2,7]. The Main focusing things of our technique is that, by using our proposal model the “Pathao” and “Uber” owners will be able to categorize all the textual documentations into Good, Bad and Average by which it is possible to know the key point what users care about. By counting these categories of text review if the good category is more than bad category then easily it can say the users concern this service is good otherwise vice-versa. Also, its possible to find out where to develop of this service. One way described in several papers that there have been work with the text document results. Better than the star rating which is in numeric shape.

1.2 Motivation

In Research Text Documentation area is very interesting area. In this area any researcher can find the different types of smart tool for achieving solution of the prominent problem.

The most important things are that, almost there has not work on transportation system in the perspective of Bangladesh website data. This base gives us the inspiration to work on it. The most vital moment that clearly inspired us that these are almost no work had been performed primarily base on Bangladesh transport dataset.

1.3 Rationale of the Study

The current transport system of Bangladesh is highly motivated by the transport review dataset, challenges of the experimental way how they fulfill their goal by using the different learning methods. In experimental text Documentation area is a very interesting area. Here any researcher can find very smart tools for achieving their desired goal.

The most vital thing that honestly motivated us that there is almost no research has been accomplished based on Bangladeshi transport review dataset. We have taken on hand with that. We think that it will be unique experiment from others.

1.4 Research Question

We have selected some question as our research work which is being answered stepwise.

1) Why are we need research?
2) How is the review being classified by the machine?
3) How will we perform the technique?
4) What is the accuracy of generated by the algorithm?
5) Which method is accelerated and accurate absolute?

1.5 Expected Outcome

The model's performance will be tested currently that we have a constructed by way of usage of different types of machine learning algorithms. The things would have tested that how good algorithms response to our data set.

About to calculating total number of words of a review dataset and also the total number of Good, Bad and the Average reviews. The research project will deliver an awesome result of the reviews of desired transport system by using the Model. Whether it is Good Bad or Average the transport owners will be useful by the way of expecting the key points of the opinion which have been given by the users of the system

1.6 Report Layout:

Chapter 1 Discusses about our thesis motivation, Rationale of the Study, Research Question and Expected Outcome.

Chapter 2 Introduce with the Background history of the research. It additionally gives us the facts of related works this research. Challenges also are mentioned right here.

Chapter 3 Discusses approximately the technique of our research work. Details works of data mining, machine learning and NLTK technique. Here also mentioned approximately about the rate collection processes.

Chapter 4 Discuss details of the result outcome and mentioned about as and out of that's project with experiment and result.

Chapter 5 Discuss our research with future scope that can be performed and conduct the thesis.

CHEPTER 2
BACKGROUD

2.1 Introduction

Features extraction is very important process in Textual Content Documentation. Evaluating the sentiment category like Positive Negative and the Neutral is a way which is categorically based on the user's emotion from the plenty of text documents in google play store through internet.

Minqing Hu and Bing Lu presented a paper name "mining opinion feature customer review" they describe with very clearly why sentiment analysis is too much important to understand the clients or the user's opinion what they experienced from the service.

But our technique is about the machine learned through difference types of machine learning Algorithms. By this learned machine, it is capable to identify totally new text document whether it Positive, Negative and Neutral.

Now a day the use of AI to detect the emotion in the text documentation is very good way because it's the combination of several built-in function and library which has been developed by different type of programing language.

2.2 Related works

Online reviews are frequently accessed through the users to shopping, see the movies to and after taking the ride service. However, given opinion in online is a beneficial aid for sentiment evaluation therefore it can be a great source of the review which carry the user's emotion. We can accomplish the work in two methods. Some works were done with just for answering the review as good, bad and average. On the Other hand, we can apply here the Rating process. Many of the process have gone both.

We can pick out the work in two ways some work had been achieved with just for to categorizing positive, negative and the neutral and some work with only giving the star. Many of the work have done in these two principles.

Text reviews in social Medias and website are revolutionary on the basis of text documentation a lot of work has been done for sentiment analysis [5] .It is mainly important things to identify the user sentiment for giving prediction of a new review finding from online resource.

Text content review in different website and social media are interesting day by day, actually a lot of work has been done based on the text documentation for analysis of sentiment[1,6,7,11],for predicting any review its very important of identity the user's sentiment which has found from online, always human sentiment collection is crucial because of its giving us, the more inside into how the sentiment of the complete excerpt is formed it component[8].

on the other hand, there has same paper which is expresses that textual documentation, result gives the better results better then star rating which is in the numeric format [4].

Score rating from the textual content evaluation has been carried out for advancing the classification [5]. using the opinion is always appreciating to all because users are the most serious person who really clearly deals with text review through online, by the uses of those review text giving only Good, Bad and Average is also a text mining for understanding the user's sentiment [1,2,7].

But few other say that classifying a text review as Good, Bad and Average, the usage of supervised learning Algorithm there may be the tendency of Good classification, accuracy is approximately 10% higher than bad one and the Neutral.

in order to alleviate they proposal advanced Naive Bayes set of rules that could explicit the common values of the two accuracies [9].But of the cases their class level, there has we counter tradition at the end [10].

A paper was published in 2014 they offer to bring the gap between the phrase level review and the document level sentiment analysis by leveraging the result given by the user's review [11].

our research offers the first approach that putting forward or predicting of class level of review is not simply enough and so we extended this through giving the answer of total of Good, Bad and the Average review of the particular transportation system and the application.

By these the owner of the transportation system or the mobile application can be capable to know the placement of the system because of the class level prediction of the total reviews which were given by the user through the services.

2.3 Research Summary

This Research study make us known that our research is about Supervised Machine learning and it is also a Classification Model Problem.

For this, we have learned Machine learning very Deeply about Classification Model and the related Algorithms. As our dataset is all approximately textual content documentation so we studied how the Natural Language Processing is used for the process of text content Documentation.

There are many Tools and Technique which is suitable for this Area. We have also found out which technique is better for the completion of our research work. We have learned different kinds of tools to complete our research. We have got several types of tool and the technique to accomplish our research. We deeply learned about NLTK and Sci-kit learning library and popular python programming language. Moreover, we learned the Cross Validation for the better Performance Measuring and also how to import this to the Sci-Kit learn Libraries.

2.4 Scope of the Problem

The working scope of our research is absolutely very broad. We can use boosting algorithm for giving rank of these transportation system and the Mobile application which reviews are being used in our training dataset. Here, we don't know about how many topics are capable for each of system. We can solve the problem by using LDA topic modeling. As our dataset is almost very new and no work had been done yet, there is a big opportunity to find out the problems and to solve it. It can be future business prediction model or recommendation system.

2.5 Challenges

To find out the word selection and model selection is a massive venture. It's now not easy to find out. We wanted a few valid datasets this is fully correct.

Feature choice become additionally a lot crucial for this study. Algorithm use was additionally an important challenge. For this research which set of rules could be very suitable, finding this was not an easy task.

When a Beginner go for a research, basically he/she needs to go with very new things. So, keep patience for studying or working it can be a challenge.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

Text mining barely exclusive manner from numeric data due to the fact in data mining. The algorithm we cannot use the text document immediately and it also cannot provide us the numeric result.

On the opposite text files are not in a tabular layout, then we cannot find the typical function which we can set in numeric data, in our research is based on classification problem so finding the category or less, from textual contents fact is another challenging part for finishing our research.

First of all, we have pre-processed the words and transformed, the text data into the numeric form, after that we have used those data into the classification algorithms

Classification algorithm one ordinarily used for supervised learning, we have used classification city of rules because our research goal is about to predict textual review content whether is good, average or bad.

Here we have made the machine learned approximately, a particular class or target. After the machine learned by using the training dataset, our machine which was been learned will give the prediction, about the data whether the review is good, average or bad.

3.2 Research Subject and Instrumentation

Our research aim could be very unique. We are working on the premises of uses assessment or opinion of level traffic service of Bangladesh. So, it's very critical pick out the users sentimental about the traffic service. For this we have got used the way to analysis, the user's overview and used the machine learn about each word and additionally the relation of every word of a review then we built a classification model that's why our research perspective is to constantan amazing by using this review text data.

After that, we learn used NLTK (natural language toolkit) for tokenizing removing prevent or stop words and also parts of speech tagging.

But here we have normally used the sci-kit learning knowledge of library which is involves though python programming language.

Now it's becoming a totally popular and useful tool for analyzing the review data. And also solving machine learning task. The important of sci-kit learning is that we can import different types of library which include specific algorithm visualization tools.

Here in our research all machine learning procedure and statistic data visualizing were execute by sci-kit learning knowledge of library.

3.3 Data Collection Procedure

Collection data is also very hard challenging, for our research studies we have gathered traffic related data from google play store “Pathao” and “Uber” application. Google play store is kind of storage of google where any kind of mobile application is available for android. We basically use the Bangladeshi local user services review, in the service of “Pathao” and “Uber”, here Bangladeshi people gives the review about the ride and the mobile application after using their services.

For considering review and rating on that google site users must have to log in first with valid account, as we are working transport service related dataset, So we have only collected the Bangladeshi traffic related “Pathao” and “Uber” transport system and mobile application review in the Google play store..

As we are working with the online transportation Data .There has so many applications with the same of name “Pathao” and “Uber” but we selected the Actual application .There has more than ten thousand of reviews in that application. From there we have selected 500 different data for our system Training Dataset. Collection of different reviews will provide better result, and finally we have valid data review for our training data set which full of Text.

3.4 Implementation Requirement

For archiving our aim, we have maintained a few steps. those all are very related with each other it's we have dealed with textual content of data our method will be isolated. Describe Figure 3.4.1 indicating the pre-processing steps of research procedure.

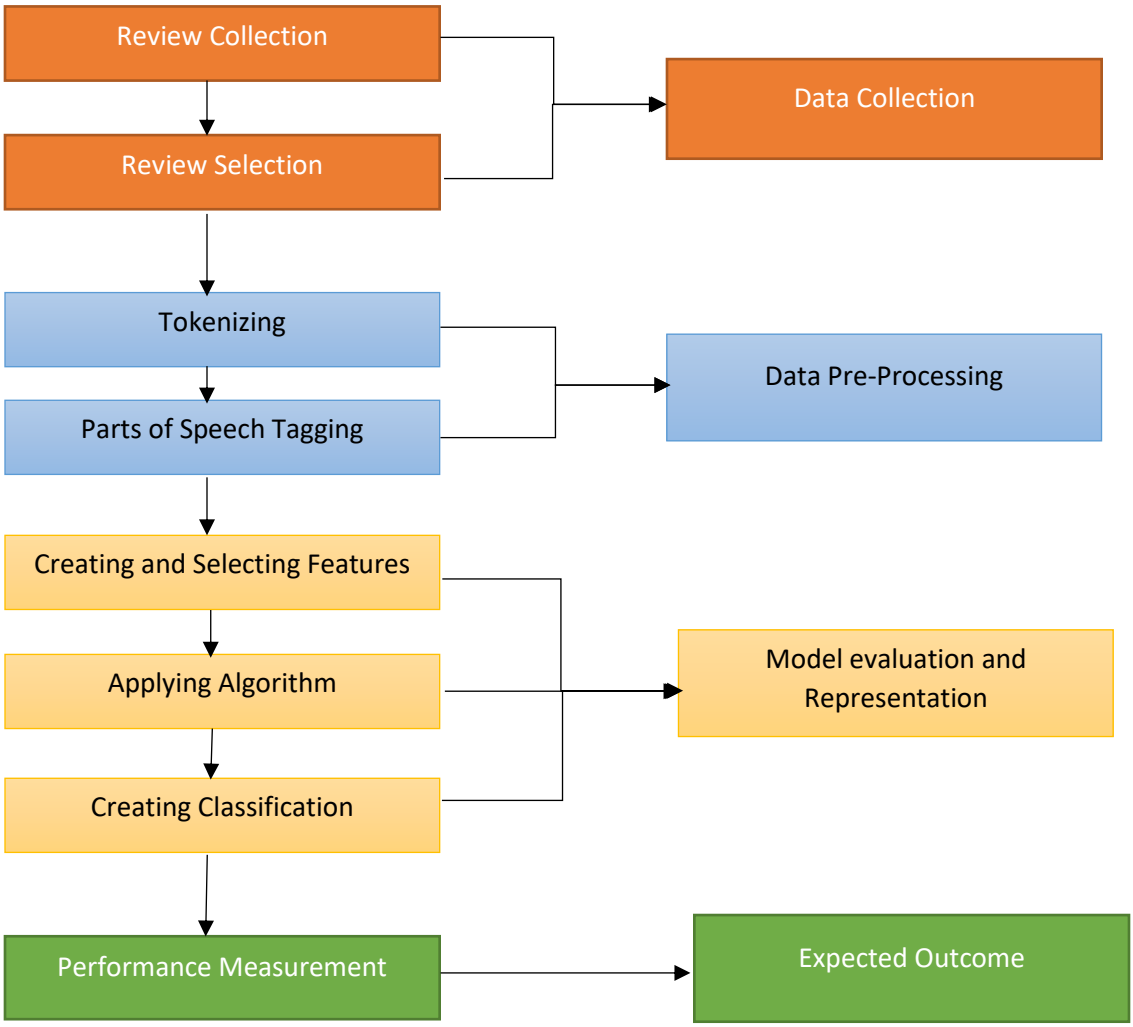

Figure 3.4.1- Research Procedure

3.4.1 Knowledge Discovery in Database

KDD process is the fundamental step for extracting knowledge from raw data. For extracting knowledge from data set this system is being use globally. For any kind of machine learning

problem, we also observed KDD process to complete our work. There have some steps of knowledge discovery from dataset:

Data cleaning: the noise and inconsistent is eliminated.

Data integration: more than one information assets are blended.

Data selection: fact applicable to the analysis are retrieved from the dataset.

Data transformation: information is converted or consolidated into from suitable for mining by means of appearing praises or aggregation operation data mining

Data mining: smart techniques are applied so one can extract fact patterns

Pattern evaluation: information styles are evaluated.

Knowledge presentation: knowledge representation steps the following diagram shows of knowledge discovery. Describe Figure 3.4.2- KDD process is used here for basic preprocessing

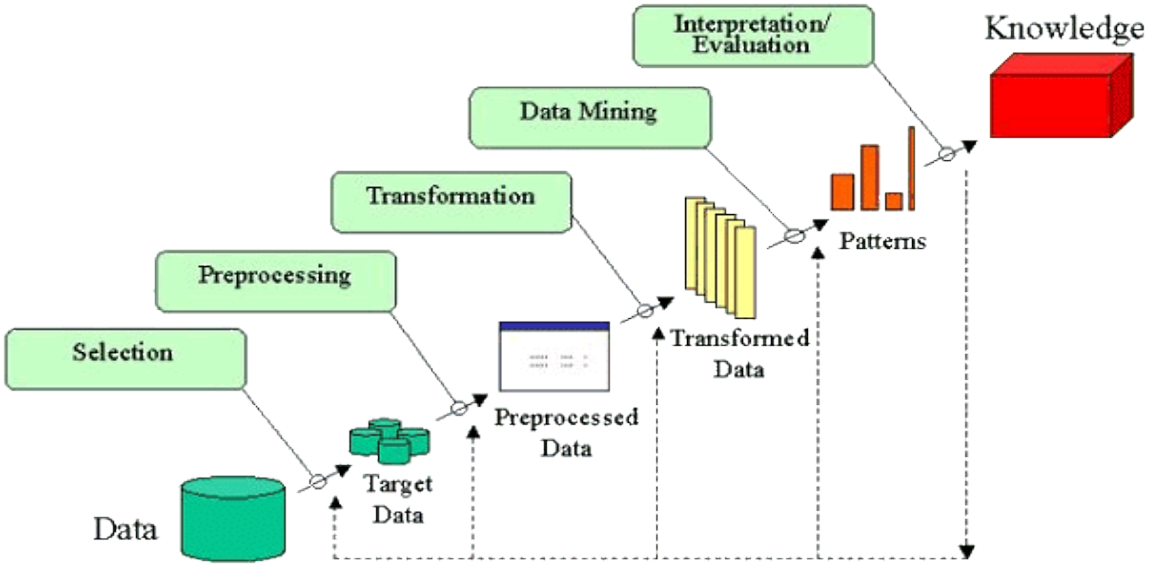

Figure 3.4.2- KDD process is used here for basic preprocessing

3.4.2 Data pre-processing:

We have selected the review for the Training dataset on the basis of the quality of the data. The "Pathao" and "Uber" application in google play store collects the users or client review. we found different types of review. Some of the review were in Bangla Motive But written with the English

word which aren't meaningful like "pathao apps ta khub e chomotkar.app load hote khub besi somoi ney na".

Generally, we avoid this type of data which creates confusion to the system. So, in our training dataset we didn't use this type of data.

As we have use sci-kit learning library, the tool is unable to identify this type of review which is not fully written in English. The Review what the users consider in the mixed language, we have remove also.

3.4.3 Tokenizing:

For understanding the meaning of the sentence in our dataset we have used sentence tokenizer. To identify total amount of word quantity we used here word tokenizer.

By using natural Language Tool-Kit

"This concept of Pathao is awesome. It really helps and saves a lot of time .In previous I used to reach my house in Shamoily by 2 hours from Tejgaon . Now it just take only 30 minutes to reach my home .Thank you Pathao .You saved a lot of time Sometimes it takes time to find out rider Developers need" Describe Figure 3.4.3.1: Sk-learn is used to tokenize

After Sentence tokenizing its look like:

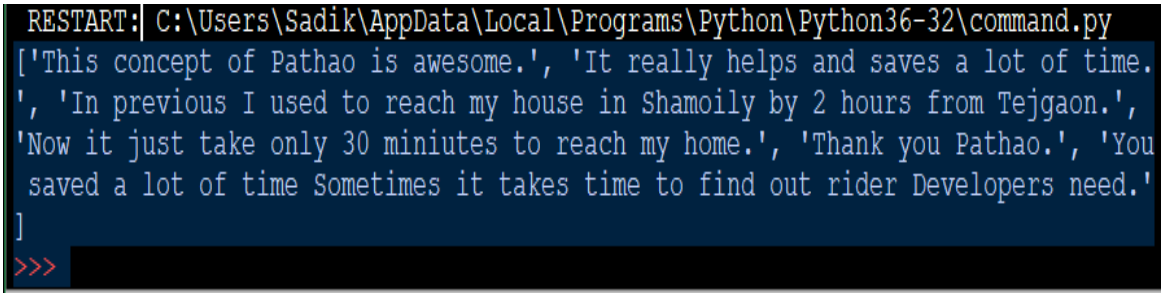

Figure 3.4.3.1: Sk-learn is used to tokenize

Describe Figure 3.4.3.2: Sk-learn is used to tokenize a word

After Word tokenizing its look like:

```
 RESTART: C:\Users\Sadik\AppData\Local\Programs\Python\Python36-32\command.py
['This', 'concept', 'of', 'Pathao', 'is', 'awesome', '.', 'It', 'really', 'helps
', 'and', 'saves', 'a', 'lot', 'of', 'time', '.', 'In', 'previous', 'I', 'used',
 'to', 'reach', 'my', 'house', 'in', 'Shamoily', 'by', '2', 'hours', 'from', 'Te
jgaon', '.', 'Now', 'it', 'just', 'take', 'only', '30', 'miniutes', 'to', 'reach
', 'my', 'home', '.', 'Thank', 'you', 'Pathao', '.', 'You', 'saved', 'a', 'lot',
 'of', 'time', 'Sometimes', 'it', 'takes', 'time', 'to', 'find', 'out', 'rider',
 'Developers', 'need', '.']
>>>
```

Figure 3.4.3.2: Sk-learn is used to tokenize a word

Here we have used natural language Tool-Kit and the specialty of NLTK is to separate is sentence and the word from each other. we get know that (.) full stop separates a single sentence from other sentence of this text documents. it also capable to identify capital letter which is starting of a new sentence.

if the sentence looks like, "Last night during the raining I Suggested Mr. Sadiik Samee hired a ride from Uber". Describe Figure 3.4.3.3: Sk-learn is used to tokenize

```
command.py - C:\Users\Sadik\AppData\Local\Programs\Python\Python36-32\command.py (3.6.3)
File  Edit  Format  Run  Options  Window  Help
from nltk.tokenize import word_tokenize, sent_tokenize


example_text = "Last night during the raining I Suggested Mr. Sadiik Samee hired a ride from Uber"
print (sent_tokenize (example_text))
print (word_tokenize (example_text))
```

Figure 3.4.3.3: Sk-learn is used to tokenize

Describe Figure 3.4.3.4: Sk-learn is used to tokenize a sentence

After Sentence tokenizing it's look like:

```
 RESTART: C:\Users\Sadik\AppData\Local\Programs\Python\Python36-32\command.py
['Last night during the raining I Suggested Mr. Sadiik Samee hired a ride from Uber']
['Last', 'night', 'during', 'the', 'raining', 'I', 'Suggested', 'Mr.', 'Sadiik', 'Sam
ee', 'hired', 'a', 'ride', 'from', 'Uber']
>>>
```

Figure 3.4.3.4: Sk-learn is used to tokenize a sentence

Here, in this Sentence full-stop (.) is not the end of sentence. NLTK is sincere about a sentence finishing (.) and which is not always. It also can identify the sentence which started with capital letter only.

3.4.4 Part of speech tagging:

In data pre-processing, part of speech tagging is one of the important feature. By using this feature, it is possible to point out of each word of sentence which parts of speech it is. In this part of the report layout we have shown an example of a review which has been represented as parts of speech tagging. Describe Figure 3.5.7: Sk-learn is used to Parts of Speech Tagging

```
>>> pos_tag(word_tokenize("“This concept of Pathao is awesome. It really helps a
nd saves a lot of time .In previous I used to reach my house in Shamoily by 2 ho
urs from Tejgaon . Now it just take only 30 miniutes to reach my home .Thank you
 Pathao .You saved a lot of time Sometimes it takes time to find out rider Devel
opers need”"))
[('“', 'NN'), ('This', 'DT'), ('concept', 'NN'), ('of', 'IN'), ('Pathao', 'NNP')
, ('is', 'VBZ'), ('awesome', 'JJ'), ('.', '.'), ('It', 'PRP'), ('really', 'RB'),
 ('helps', 'VBZ'), ('and', 'CC'), ('saves', 'VBZ'), ('a', 'DT'), ('lot', 'NN'),
('of', 'IN'), ('time', 'NN'), ('.In', 'NNP'), ('previous', 'JJ'), ('I', 'PRP'),
('used', 'VBD'), ('to', 'TO'), ('reach', 'VB'), ('my', 'PRP$'), ('house', 'NN'),
 ('in', 'IN'), ('Shamoily', 'NNP'), ('by', 'IN'), ('2', 'CD'), ('hours', 'NNS'),
 ('from', 'IN'), ('Tejgaon', 'NNP'), ('.', '.'), ('Now', 'RB'), ('it', 'PRP'), (
'just', 'RB'), ('take', 'VB'), ('only', 'RB'), ('30', 'CD'), ('miniutes', 'NNS')
, ('to', 'TO'), ('reach', 'VB'), ('my', 'PRP$'), ('home', 'NN'), ('.Thank', 'NN'
), ('you', 'PRP'), ('Pathao', 'VBP'), ('.You', 'RB'), ('saved', 'VBN'), ('a', 'D
T'), ('lot', 'NN'), ('of', 'IN'), ('time', 'NN'), ('Sometimes', 'NNP'), ('it', '
PRP'), ('takes', 'VBZ'), ('time', 'NN'), ('to', 'TO'), ('find', 'VB'), ('out', '
RP'), ('rider', 'NN'), ('Developers', 'NNS'), ('need', 'VBP'), ('”', 'NNS')]
>>>
```

Figure 3.4.4: Sk-learn is used to Parts of Speech Tagging

In a sentence for finding the Good, Bad and Neutral identify keyword we basically focus in the Adjective (tagged as ’JJ’) and Name (tagged as 'NN') because we knew that other words not content any kinds logical information or not mean any sentence of user. Here in Table **3.4.4.1** we shown Adjective words list what has used in Our Dataset.

Table 3.4.4.1- Showing Adjective Word for our Dataset

Number	Positive Words	Negative Words	Neutral Words
1	Good	Worst	Daily
2	Awesome	Poor	Dependable
3	Interesting	Pathetic	Immediate

4	Enjoyable	Problem	Commercial
5	Great	Bad	Communication
6	Excellent	Suffer	Internal

3.4.5 Feature Extraction

Feature extraction from the pre-processing dataset is a very important part, to know the function of the algorithms it helps a lot. In the sci-kit learn Library a different type of technology is being used for feature extraction, it helps us to extract a feature from our textual content document. For this it has no longer be used inside the model, it makes use of the pattern that import it, instantiate it, fit, transform and predict.

We have used 408 reviews for our training and testing data set. Here a bag of words have been used for extracting feature from review. These words are a sparse vector of occurrence counts of words. We have found 1164 words in our full data set and the sci-kit learning library consider each word as a feature. That's why there are 1164 features. As our goal is detecting Good, Bad, Average review, here our total classes are just three in number. So, we have created 3D vector. 3*1164 represents the class and feature number. In figure, the last 17 review's finding feature position and extraction which is being evaluated by sci-kit learn library is given below figure 3.4.5

391	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
392	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
393	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
394	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
395	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
396	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
397	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
398	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
399	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
400	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
401	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
402	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
403	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
404	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
405	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
406	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
407	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

408 rows × 1164 columns

Figure 3.4.5 shows the feature position of last ten reviews

3.4.6 Applying Algorithm:

Our research is preliminary based on supervised learning. Multi classification is the supervised learning model. So, we have used here iteration algorithms there are different types of the tools available for using the algorithm and the procedure are different, in our sci-kit learn library become it is quite different from other.

Different algorithms show also different results. So, selecting the proper algorithm is also a several tasks. Here we have selected Logistic Regression Algorithm, Support Vector machine and K-neighbors classifier. By using these algorithms, we have used a classifier model.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

For gaining a good result, our research project depends on the selection of appropriate Data and the perfect Model. For Building a Model by way of the use of textual content Documentation and finding proper result the language processing Part is very necessary.

4.2 Experimental Results

To achieving the expected goal, we have used here three different classification Algorithm. The Logistic Regression Algorithm gives us the best accuracy which is around **79.61%**. In Table 4.2.1 represents the accuracy of all Model Performance.

Table 4.2.1- Performance Measurement Accuracy using different algorithms

Number	Algorithms	Result
1	Support Vector Machine Algorithm	**78.64%**
2	Logistic Regression Algorithm	**79.61%**
3	K Nearest Neighbor Algorithm	**77.45%**

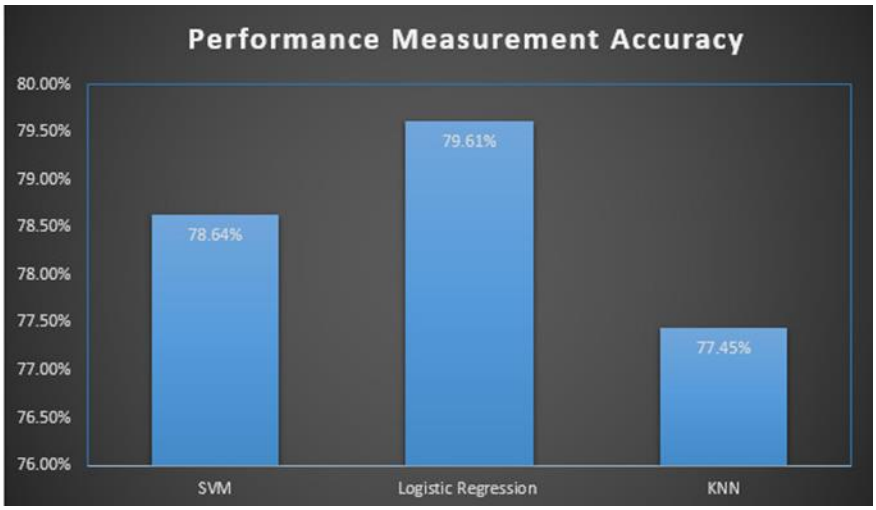

Figure 4.2.2 shown three different Models results by using those three algorithms.

4.3 Descriptive Analysis

For measuring the overall performance, we have used cross validation. Cross validation is such type of technique by cross validation it's possible to measure any kinds of prediction model and also it can validate the result. Here, in every time of its measurement process. As our goal is to predict the class whether it is Good, Bad or Average that's why cross validation is much better effective way.

Table 4.3.1- shows the review text and the result

No	Review/Text	Result
1	'Pathao apps is Good And It's Ride Is also best ride in Dhaka City'	1
2	'Pathao service is worst day by day and their car is also need repair'	0
3	'Road was fine that why car running smoothly'	2
4	'Excellent service provider and also Driver Behavior is good'	1
5	'The service is good enough but the app is disgusting'	2
6	'Very bad maintaining They only see their profit'	0

7	‘Awesome Ride by Pathao and with their smart driver’	1
8	‘Negative evaluation for ultimate update suffering greatly’	0
9	‘Great service real time money saver gives peace of mind’	1
10	‘Made life more easy Thanks Pathao team Good luck’	1

Here we have also shown the percentage of total number of Good, Bad and Average Review Which Our Prediction Model Predicted.

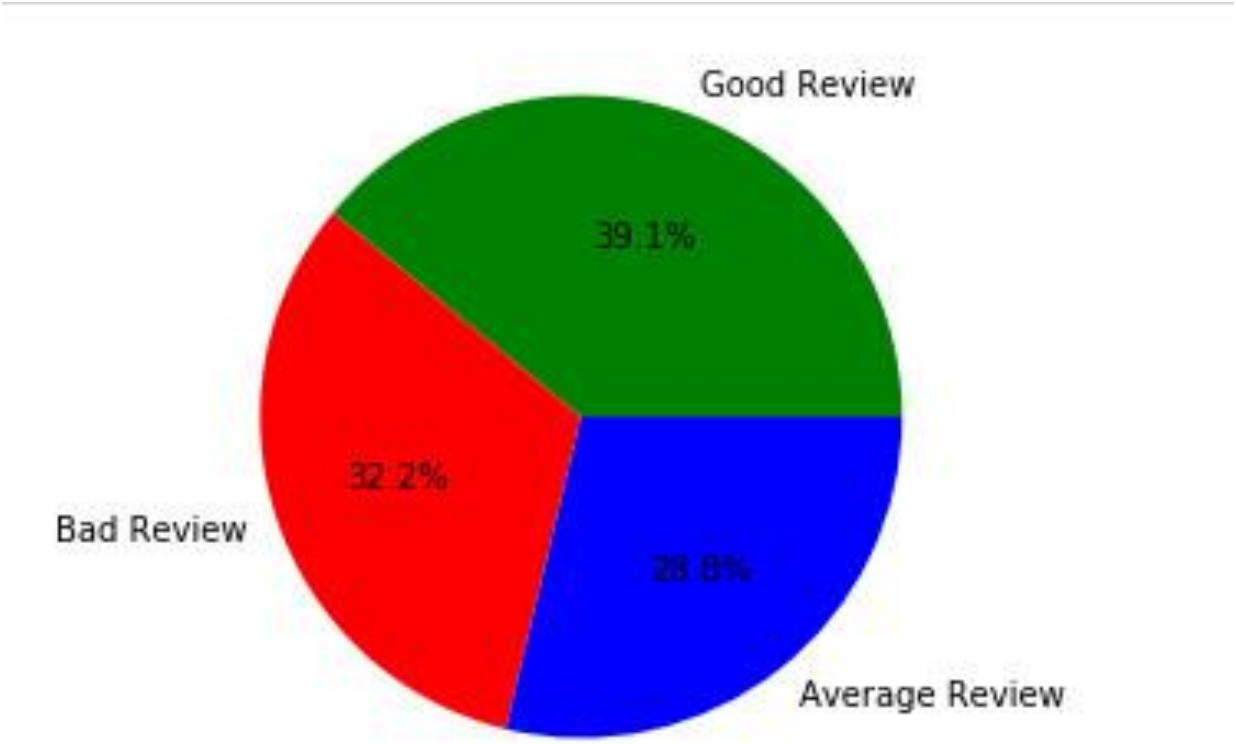

Figure 4.3.1: Pie chart shows the percentage of Machine Predicted Review

4.4 Summary

After preprocessing the dataset a model has been created using Multinomial Naïve Bayes and it will predict the answer. Describe figure **4.4.1** finding prediction is shown.

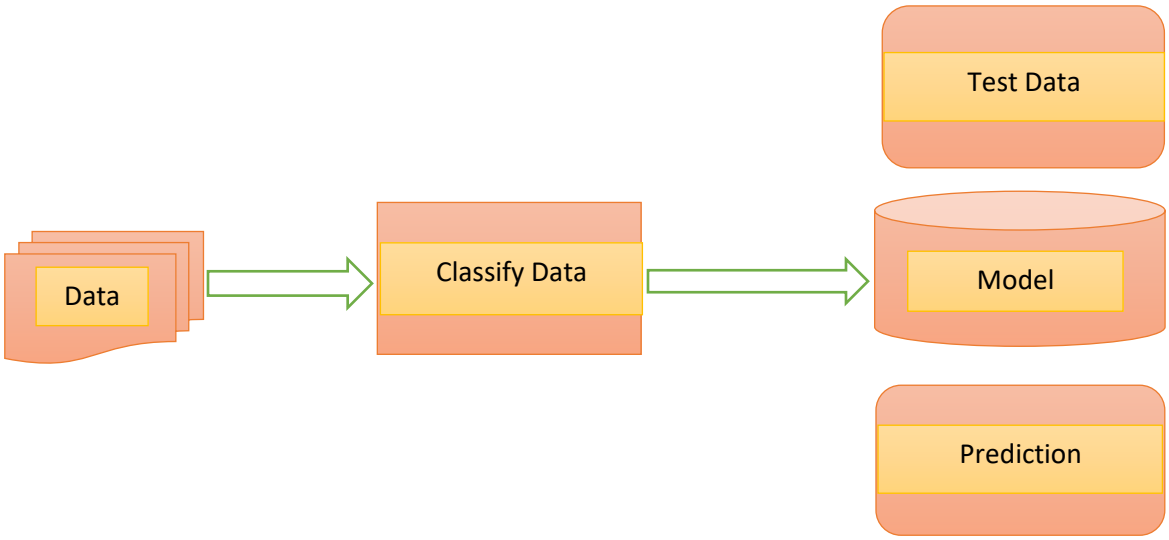

Figure 4.4.1: whole process of finding prediction

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

In This part of our Report Layout we have described about some important aspects of our research which are the brief summary of research, recommendation, implication of the research work and finally the Future Scope of this Research. Actually, part by part all the portion has been described below:

5.1 Summary of the Study

To complete our Research Work, we have studied how the Machine learning techniques are used for getting the solution of various types of machine learning problem. There are two kinds of Machine learning process that is Supervised Machine Learning and Unsupervised Machine learning. This Study make us known that our research is about Supervised Machine learning and it is also a Classification Model Problem.

For this, we have learned Machine learning very Deeply about Classification Model and the related Algorithms. As our dataset is all approximately textual content documentation so we studied how the Natural Language Processing is used for the process of text content Documentation.

There are many Tools and Technique which is suitable for this Area. We have also found out which technique is better for the completion of our research work. We have learned different kinds of tools to complete our research. We have got several types of tool and the technique to accomplish our research. We deeply learned about NLTK and Sci-kit learning library and popular python programming language. Moreover, we learned the Cross Validation for the better Performance Measuring and also how to import this to the Sci-Kit learn Libraries.

5.2 Conclusion

For the Beginners, Machine Learning is very good research area. In the basis of our Country related work like user's opinion classify form traffic dataset has not yet been done. We tried to establish

a model which is able to predict the textual content data which is consider by a user. By this, the system owner can take some powerful step in future for the betterment in their services. Basically, we used all the review form Online which has been considering by the service taker. This review is only that expression what they received form the services. The Pointing things is that all of the data which we already collect are live because with respect of the time the people are taking the services and considering the review what they feel about the service. We tried to describe all the operating techniques, Working procedure, models and also the method with table and Figure. We have a plan to make this research finished in extra requirements additionally. We have additionally confronted loads hassle to complete this research. There became so noisy facts. All the Procedures have been tremendous in sum so it took a lot of time to understand and put in force this in our studies. For collecting the data from the valid source, we went with the hassle. We have experienced a few different problem that were in the beginning of our research.

5.3 Recommendations

Though we mention some related work but its amount is very few and directly there has no work like us based on Bangladeshi Data. We understand all their research process and work style after that we started to fix our research goal. After a hard effort by doing all of the work step by step finally, we are at a stage what can be said it is our expected research goal. So, for making this kind of research work it need a tremendous work for guiding us through the right path of research. We have experienced some different problems that were inside the starting of our research. We have also stuck with the mastering of the large field of Data mining and Machine Learning. With the total journey if this research Work our supervisor Ms. Nasrin Akhter madam helped us a lot and guided us for making this research project successful.

5.4 Implication for Further Study

From this research work it could be finished LDA topic modeling by using which transport system owners can become aware of what the customers focus on specially. We will be able to rank the system by means of the use of boosting algorithm. Also it will be possible to give the exact or real ratings to the transport System and the Mobile Application both.

The Another one thing can be point as the future work or study, that is it will be possible to give the ratings mark among the same types of the Websites Based on some Particular topic. For this it need a **Web crawler system** to collect data from different websites based on particular object.

After collecting that data, it will be stored in a database. Then the further process is to modification that data by using our proposed model and make the machine learned it's possible to categories all the data. Then compare it to other websites data it's possible to make a sorted of rank. And hope so users will be capable to use these system by using Android Mobile Application.

REFERENCES

[1] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[2] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[3] Ghose, Anindya, and Panagiotis G. Ipeirotis. "Designing novel review ranking systems: predicting the usefulness and impact of reviews." Proceedings of the ninth international conference on Electronic commerce. ACM, 2007.

[4] Ganu, Gayatree, Noemie Elhadad, and Amélie Marian. "Beyond the Stars: Improving Rating Predictions using Review Text Content." WebDB. Vol. 9. 2009.

[5] Lee, Moontae, and R. Grafe. "Multiclass sentiment analysis with restaurant reviews." Final Projects from CS N 224 (2010).

[6] Socher, Richard, et al. "Semi-supervised recursive autoencoders for predicting sentiment distributions." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011.

[7] Zhang, Ziqiong, et al. "Sentiment classification of Internet restaurant reviews written in Cantonese." Expert Systems with Applications 38.6 (2011): 7674-7682.

[8] Wu, Jean Y., and Yuanyuan Pao. "Predicting Sentiment from Rotten Tomatoes Movie Reviews."

[9] Kang, Hanhoon, Seong Joon Yoo, and Dongil Han. "Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews." Expert Systems with Applications 39.5 (2012): 6000-6010.

[10] Khairnar, Jayashri, and Mayura Kinikar. "Machine learning algorithms for opinion mining and sentiment classification." International Journal of Scientific and Research Publications 3.6 (2013): 1-6.

[11] Zhang, Yongfeng, et al. "Do users rate or review?: Boost phrase-level sentiment labeling with review-level sentiment classification." Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014. ©Daffodil International University

[12] Thelwall, Mike, et al. "Sentiment strength detection in short informal text." Journal of the Association for Information Science and Technology 61.12 (2010): 2544-2558.

Plagiarism Checker Screenshot