

**Spam Text Detection in Bangla language and Comparison Between Machine
Learning Algorithms**

BY

Shovon Ahammed

ID: 162-15-7671

AND

Md. Mostafizur Rahman

ID: 162-15-7764

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

Supervised By

Dr. Md Ismail Jabiullah

Professor

Department of CSE

Daffodil International University

Co-Supervised By

Mr. Ahmed Al Marouf

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY 2020

APPROVAL

This Project titled “**Bangla Spam Text Detection and Comparison Between Machine Learning Algorithms**”, submitted by **Shovon Ahammed**, ID No: **162-15-7671** and **Mostafizur Rahman** ID No: **162-15-7764** 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 **07-08-2020**.

BOARD OF EXAMINERS



Dr. Syed Akhter Hossain
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

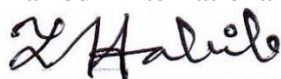
Chairman



Dr. Fizar Ahmed
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Tarek Habib
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Mohammad Shorif Uddin
Professor

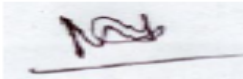
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Dr. Md Ismail Jabiullah, Professor, 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.

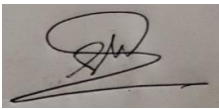
Supervised by:



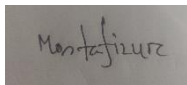
Dr. Md Ismail Jabiullah
Professor
Department of CSE
Daffodil International University
Co-Supervised by:



Mr. Ahmed Al Marouf
Lecturer
Department of CSE
Daffodil International University
Submitted by:



Shovon Ahammed
ID: 162-15-7671
Department of CSE
Daffodil International University



Md. Mostafizur Rahman
ID: 162-15-7764
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 **Dr. Md Ismail Jabiullah, Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

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

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

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

ABSTRACT

Due to the advancement of technology electronic communication has reached its peak. It has great impact on our life. But some dishonest people are using it to fulfil their evil wish. Millions of people use Bangla text in internet. They communicate in e-mail, Facebook, twitter and other social networking platforms in Bangla language. And some people are manipulating this platform. They are sending spam messages to the users. Many works have been done for spam mail in English language. But there is no significant work for spam messages in Bangla language. This inspired us to work for spam message detection in Bangla language. Our approach to detect spam messages in Bangla language. For this approach 6 types of spam messages have been considered. We made the dataset for the task. We have collected the data from different platforms. Then we have used Natural Language processing to perform spam mail detection. We have compared between traditional NLP algorithm. For support vector machine we have got an accuracy of 96%.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
CHAPTER	
CHAPTER 1: Introduction	1-4
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the study	3
1.4 Research Questions	3
1.5 Expected Output	4
1.6 Report Layout	4
CHAPTER 2: Background	5-6
2.1 Preliminaries	5
2.2 Related Works	5
2.3 Comparative Analysis and Summary	6
2.4 Scope of the problem	6
2.5 Challenges	6

CHAPTER 3: Research Methodology	7-15
3.1 Research Subject and Instrumentation	7
3.2 Data Collection Procedure	7
3.3 Statistical Analysis	9
3.4 Proposed Methodology	11
3.5 Implementation Requirements	15
CHAPTER 4: Experimental Results and Discussions	16-27
4.1 Experimental Setup	16
4.2 Experimental Result & Analysis	16
4.3 Discussion	26
CHAPTER 5: Impact on Society & Environment	28
5.1 Impact on Society	28
5.2 Impact on Environment	28
5.3 Ethical Aspects	28
CHAPTER 6: Summary, Conclusion, Recommendation and Implication for Future Research	29
6.1 Summary of the study	29
6.2 Conclusions	29
6.3 Implication for further study	29
REFERENCES	30-31
PLAGIARISM REPORT	32-33

LIST OF FIGURES

FIGURES	PAGE NO
Figure 1.2.1: Report on BBC about Bangla spam message	2
Figure 3.2.1: Dataset after annotation	8
Figure 3.3.1: Box plot of dataset	9
Figure 3.3.2: Phrase length of ham data	10
Figure 3.3.3: Phrase length of spam data	11
Figure 3.4.1: Visualization of data without preprocessing	12
Figure 3.4.2: Flow diagram of proposed approach	14
Figure 4.2.1: Accuracy of algorithms	17
Figure 4.2.2: Confusion matrix of svm	18
Figure 4.2.3: AUC-ROC curve of svm	19
Figure 4.2.4: Confusion matrix of Naïve Bayes	20
Figure 4.2.5: AUC-ROC curve of Naïve Bayes	21
Figure 4.2.6: Confusion matrix of Logistic Regression	22
Figure 4.2.7: AUC-ROC curve of Logistic Regression	23
Figure 4.2.8: Confusion matrix of Decision Tree	24
Figure 4.2.9: AUC-ROC curve of Decision Tree	25
Figure 4.3.1: Performance of algorithms on ham data	26
Figure 4.3.2: Performance of algorithms on spam data	27

LIST OF TABLES

TABLES	PAGE NO
Table 3.1.1: Dataset distribution of ham and spam	7
Table 4.2.1: Classification report of Support Vector Machine	19
Table 4.2.2: Classification report of Naïve Bayes	20
Table 4.2.3: Classification report of Logistic Regression	22
Table 4.2.4: Classification report of Decision Tree	25

CHAPTER 1

Introduction

1.1 Introduction

Spam message is a great problem for internet users. The message which are meant to harm the receiver or does not have any meaning is called spam message. The spam message can contain message about lottery winning, adult content and false scholarship. Sometimes the spam messages contain links of different harmful website. They are phishing websites. When the receiver clicks on those links, the user losses valuable information to anonymous people. People of different parts of the world are now connected to each other with the help of the internet. Now people can communicate with anyone in the blink of an eye. At the same time, dishonest people can harm users by abusing this technology. Bangla ranked at 7th among most spoken languages worldwide [1]. This represents the popularity of Bangla language worldwide. It shows Bangla language is not only confined in Bangladesh but also throughout the world. Bangla has come from Sanskrit. This language has age-old tradition and culture. Almost 230 million people used Bangla as the first language and 37 million people use Bangla as the second language [2]. And the number is going up every year. But it is a matter of great sorrow that there are many cases of spam messaging in the Bangla language. As the online transaction have increased to a great extent, the spam message in Bangla language is matter of great concern. Many mobile financial services are very popular in Bangladesh. Among them bkaash, nogod, rocket, dutch bangla mobile banking are most famous ones. Everyday millions of transactions take place on these platforms. Cybercriminals are targeting users of these platforms by sending different types of spam text in Bangla.

Now some cybercriminals are using hoax to attract the user. They advertise their offer in different platforms. Sometimes they boost their Facebook pages targeting specific people. And those pages are full of link of phishing sites. Now a day's spammers are targeting mass people by using adult content. They embed the adult content with phishing sites. When the user clicks on the link, the link takes the user to phishing site. Where the user can lose his valuable information like banking data.

Again, some cybercriminals are targeting students during different important examination. Through different social networking platforms, they reach the students. They offer the student to supply question of different examination. Then they ask for money before exam and when the student with less determination give the money, they block the student. This way they are breaking the confidence of the student. Sometimes they supply authentic question to the student. These cybercriminals are injuring our education system very badly.

1.2 Motivation

Bangla is the state language of Bangladesh along that millions of people speak in Bangla as their first language. They use different social networking platforms in Bangla language. Moreover, mobile banking is booming in Bangladesh. So, the dishonest people are targeting innocent people to get their valuable information. According to BTRC, the total number of internet subscriber has reached 99.984 million at the end of February, 2020[16].

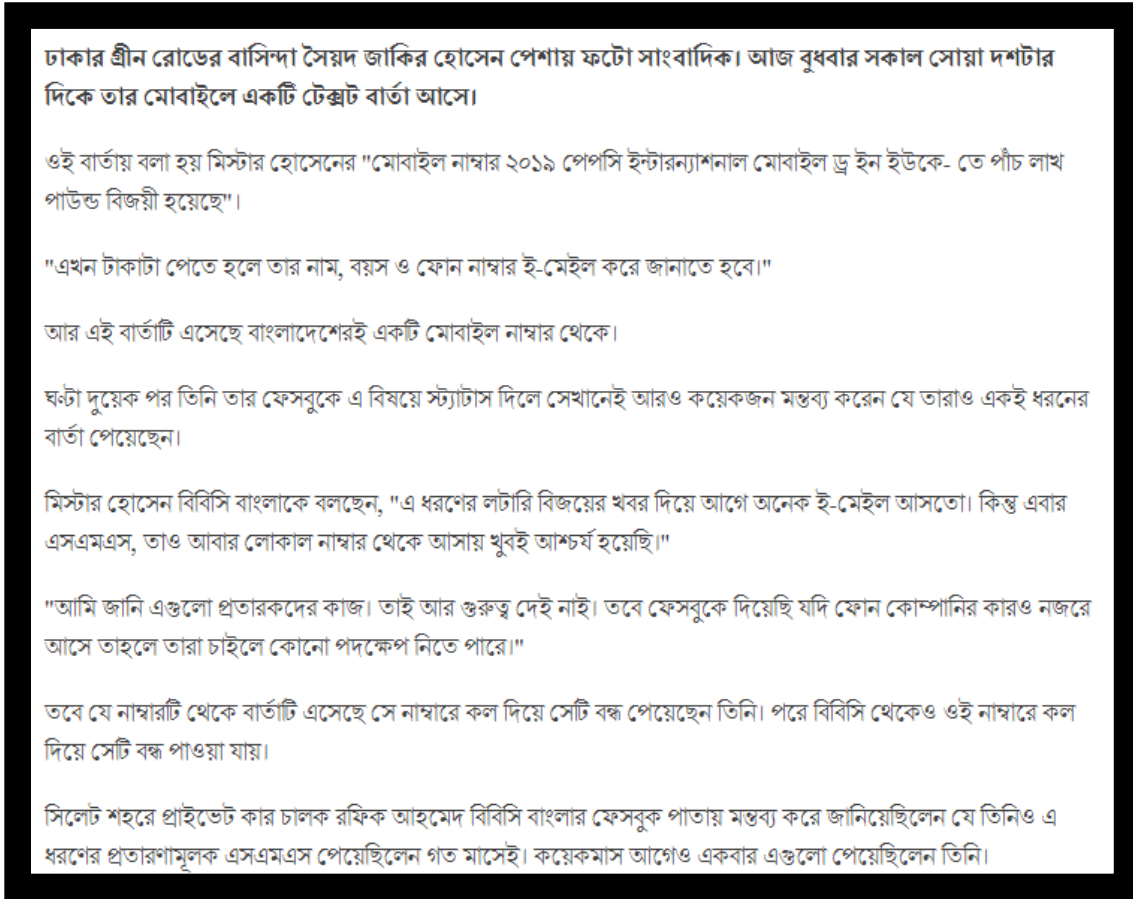


Figure 1.2.1: A report on BBC about Bangla spam message

In Figure 1.2.1 we can see statement of a victim [3]. There are many cases of Bangla spam message now a days. It is a matter of great concern. One day one of my close friends received Bangla spam message in Facebook. The message was about a phishing site. The message was about free tip to cox's bazar. So, to get the offer the message instructed him to register on the phishing site. As, he registered to the phishing site. After some moment he was unable to login into his Facebook account.

There are many works for spam messages or mail in English language. But there is no significant on Bangla spam message.

1.3 Rationale of the study

The number of internet user in Bangladesh has been increased to a great extent in last 5 years. According to a report of world bank in 2017 states that almost 14% people Bangladesh uses internet [4]. The data are getting cheaper and people are on the internet more. People in general now uses Facebook, WhatsApp, Imo. Most of the user uses this platform on Bangla language. About 80 percent internet users of Bangladesh are on Facebook [5]. Spammers are targeting people using Bangla messages. During different board examination, the cyber criminals tries to reach the examinees. They try to entice the examinees by saying that they have leaked question of board exam. Then they demand money to the examinees for questions. This way every year criminals destroy the career of many students. Some spam text contains advertisement of harmful products. They entice people by giving false hope. Some spam text contains adult content. They embed link of harmful website with adult content. When the user clicks the link, it takes them to harmful websites. We have observed that most of the spam text are money related. Moreover, no dataset related to spam message in Bangla language is available. So, we had decided to work with Bangla spam text.

1.4 Research Questions

Our approach is to detect Bangla spam text. So, there were a few questions on performing our approach:

1. What is spam message?
2. What is the difference between Bangla grammar and English grammar?
3. How to collect data?
4. How to preprocess them?
5. How to use the data to detect Bangla spam text?
6. What are most popular techniques to detect spam text?
7. Which machine learning algorithms will perform good?
8. How to compare performance of machine learning algorithms?

1.5 Expected Output

Our work is about spam message detection and comparing different machine learning, deep learning algorithm to find the best accuracy.

1.6 Report Layout

This report is divided into six sections. The very first section is introduction. This introduction section has got six part. In the first this paper describes about the introduction of our work. This part contains a brief introduction of our work. The next part is motivation. In the motivation part we have described our motivation behind the work. Our main motivation was that there is no significant work of Bangla spam text. Then we have described about rationale of the study. The next part is research questions. This part is very important. Because our approach hugely depends on this. We have given our research questions on this part. On the expected output part, we have shown what will be our output when the approach accomplished. This will help the reviewer to understand our work.

The second section is background. The background has got 5 part. The first part is preliminaries. In that we have described about different types of spam text. We have worked with six types of Bangla spam message and we have described them on that part. The next section is related work. This section has played very significant impact on our work. Though there are no work of Bangla spam text. But there are work of spam in other languages. We have reviewed those on this part.

The third section is research methodology. It has got the sub-section. In this section we have consulted our approach. We have discussed about data collection, statistical analysis and implementation.

The fourth section is experimental result and discussion section. This one is the most important section. In this section we have illustrated the performance of different algorithms through accuracy, precision, recall and f-1 score. Then we have done comparative analysis between the algorithms.

The fifth section is impact on society and environment. In this section we have discussed about the impact of our approach.

The sixth and the last section is summary section. In this section we have concluded our work and gave some recommendation.

CHAPTER 2

Background

2.1 Preliminaries

We have considered six types of spam messages.

Adult content are website links of porn sites. When the user clicks on those links it takes them to porn sites or some harmful websites that forces user to install unwanted harmful applications.

Lottery is one of the most famous and old spam message type. The spammer entices the receiver by saying that he has won lottery.

Harmful Ads are ads of those types of products that contains harmful chemical.

Educational spam messages are great problem now a days. The spammers target different board exams. They spread rumor of question leaking. They tell the student they would provide the student leaked question and they ask for advance money.

Hoaxes message are consisting of miracle offer or promises. Mostly they target Facebook users. They offer out of the world discount or job.

Money Scams is the main motive of spam messages. But there are some spam messages which targets credit card information, mobile banking information.

We have used web scrapper to scrap the data from Facebook. We have labeled our data in two categories either it is a spam or ham. The contribution of our work is to creating a new dataset for spam message detection in Bangla Language and applying machine learning algorithm to detect it.

2.2 Related Works

There are several methods for spam and ham mail classification. Each of the method have advantage and disadvantage. Anuj Kumar Singh et.al [6] have worked on different machine learning techniques. They have worked with feature selection and without feature selection. Their approach with feature selection was more accurate. They have used Fuzzy C Means in their approach. They have suggested a method using Fuzzy which have a threshold of 0.5. Neural network has been used by Alexey S. Katasev et al. [7] to create spam filter. Nikhil V Mathew et al. [8] have showed the effectiveness of n-gram techniques for spam filtering. With 5-grams they got an accuracy of 91.48%. They have used Naive Bayes classifier with n-grams. But analyzing different parameters they showed that 4-grams are more effective. Manmohan Singh et al [9] have used Non-Linear based Support Vector Machine based classifier. They have showed comparison between Linear kernel and Gaussian kernel. Linear kernel got maximum accuracy of 99.8% and Gaussian kernel got maximum accuracy of 99.9%. In their approach Linear kernel got the best result. Part of Speech Tagging with K-means algorithm have been used by Mohammad Reza Parsaei et al and they got precision of 83%. They have used Gate software to label e-mails. Ms T. Indhumathil et al. [10] have shown an efficient scheme for identifying spam bots and

terminate mailing. As spam mails waste bandwidth and storage. They have used bloom filers to handle large amount of data. Pingchuan Liu et al. [11] have presented content-based spam email filtering technique. Their experiment got an accuracy of 92.8%. Prajakta Patil et al. [12] have used svm and obfuscation URL detection algorithm to detect spam and phishing mails. Mohammad Reza et al. [13] have used part of speech tagging to detect spam mail. By using K-Mean algorithm they got a precision of 83%. Simranjit Kaur Tuteja et al. [14] concentrated on BPNN classification algorithm for email spam filtering. Harpreet Kaur et al. [15] proposed spam email classification method using integrated particle swarm optimization and decision tree. To increase the accuracy the have used unsupervised filtering technique.

2.3 Comparative Analysis and Summary

In our approach we have detected Bangla spam text by using machine learning and we have shown comparison between the performance of machine learning. We have used popular machine learning algorithm such as Support Vector Machine, Naïve Bayes, Decision tree, Logistic regression. We have shown their precision, recall, f1 score, accuracy.

2.4 Scope of the problem

There are many works on spam text of other languages. But there is no significant work on Bangla text. Bangla language is one of the most popular language in the world. According to a report of Dhaka Tribune Bangla ranked at seventh among hundred most spoken languages worldwide [Dhaka tribune]. The internet users of Bangla language are many. So, all of these facts encouraged us to work with Bangla spam text.

2.5 Challenges

Our work is related with machine learning. Dataset played a crucial part on our work. As our work is first on Bangla spam text. There was no available dataset for Bangla spam text. So, we made the dataset of Bangla spam text. Forming the dataset was challenging part of our work. We have used web scrapper to collect the data. We have analyzed different affairs of Bangla spam message. After collecting the data next challenge was to train the data to detect Bangla spam text. As the algorithms works with number, we have used count vectorizer to transform the data. The we have used Term Frequency Inverse Document Frequency to utilize the key feature of Bangla dataset. Getting the best result from the algorithms was big challenge. Tuning different features of algorithms, we have tried to get the best output from the dataset.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

Our approach is to detect Bangla spam text. Our approach is based on machine learning. We have used four machine learning algorithms and compared their performance. To train our model data is important. Without data machine cannot learn. There is no available dataset for Bangla spam message. So, we have to made the dataset. We have used web scrapper to collect the data. Most of our data is from Facebook. Facebook is one of the most popular social sites in Bangladesh. Almost all of the internet users use Facebook. That is why we targeted Facebook to collect the data.

Table 3.1.1: Dataset distribution of ham and spam

Number of data	Ham data	Spam data
1684	1223	461

Table 3.1.1 shows the dataset distribution between ham and spam data. As the amount of spam data is very low in real life. So, the amount of spam data is very low in our dataset.

3.2 Data Collection Procedure

Our approach is to create a dataset of Bangla spam text. Then using that dataset to detect spam text in Bangla language. So, data collection is the key part of our work. There is available dataset on other language for spam text. But there is no dataset for Bangla spam text. So, we decided to make a dataset of our own. We wanted to make a dataset which will be consisted of all types of spam message. So, we analyzed different platforms to collect spam message. Then we found some Facebook pages which is full of spam texts. So, we targeted them to collect education related spam messages. We have collected hoaxes from different Facebook groups. Some spammers try to target people by posting hoaxes on many big public groups. So, they can target as many people as they can. For adult content related spam content, we have used imo, ucbrowser. These two applications are widely known for their adult content. They give notification of different types of adult content.

Data collection is very important part of our work. There is no available dataset of Bangla spam message. We formed the dataset to complete our task. As we have worked with binary classification so, our dataset was divided into two class Spam Bangla messages and Ham Bangla messages. As our approach is to detect spam Bangla messages. The system should

detect Bangla spam messages along that it should be careful about Ham messages. In case of Ham messages, the criteria of Ham message are big. But spam messages are very specific. In our dataset we have considered 6 types of spam messages. They are adult content, lottery, ads of harmful products, educational, hoaxes, money scams. Our dataset contains Bangla spam message of this categories. The rest of the dataset is full of Ham messages. To collect Ham messages, we have scrapped Facebook using web scrapper.

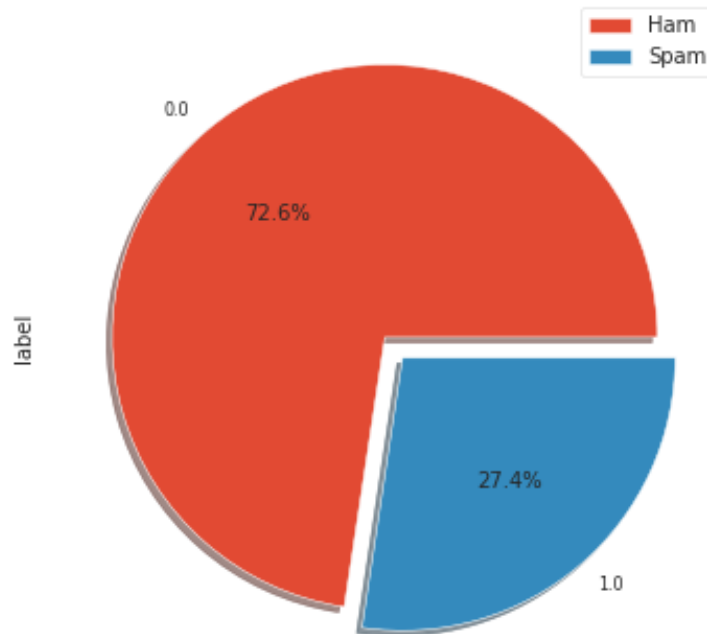


Figure 3.2.1: Dataset after annotation

In the Figure 3.2.1 we can see that almost ninety percent of the dataset are ham messages. If we compared it to our daily life, we can understand it. Spam messages is not something that we get too often. Spam message is very rare in our life. But when it comes it brings a great source of risk.

Facebook is one of the biggest social network platforms. It generates tons of data every day. People of all ages and groups use Facebook. So, we decided to take Bangla spam data from Facebook. Scraper helps us to get data from websites. We have used web scraper to get the data from Facebook. Collection of Spam data was big challenge for us. We analyzed different incident and collected spam messages from the victims. There are different public

group where people share their experience on different incident. They make people aware on different issues. So, they share different Bangla spam messages there. Many victims also shared spam messages. We have made the dataset of spam message along with regular speech. We have collected data from Facebook and SMS. Facebook is one of the most popular social networking sites in Bangladesh. Millions of people in Bangladesh use Facebook. And most of them use Bangla language on Facebook. The Internet pack is getting cheaper day by day. The number of internet users is increasing. Now people of every part of Bangladesh are using Facebook. So, the immoral peoples are getting more targets. They are targeting people through Facebook and SMS mostly. And now a days the spam message has increased to a great extent.

3.3 Statistical Analysis

Statistical analysis is used to get knowledge about the data. Every data is significant. They have their own patterns and value. For example, spam data and ham data have variations with text length. But while we have analyzed our dataset, we have found that text length of ham data has does not have any significance to find any pattern. But the text length of spam data has great significance. In our day to day life the amount of spam data is very low compared to ham data. In our approach we have worked with text data. Bangla language has very complex grammar

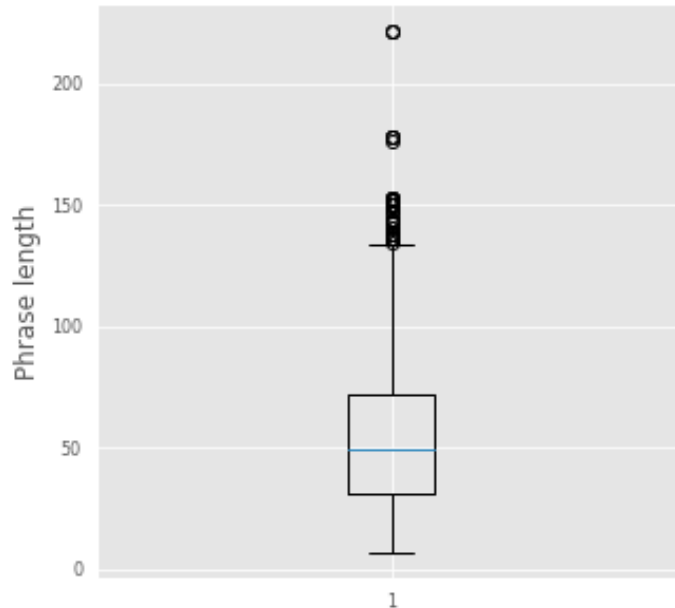


Figure 3.3.1: Box plot of dataset

In Figure 3.4.1 we can see the box plot for phrase length of the dataset. The minimum text length lies between ten to fifteen. The maximum number of phrase length lies between thirty to seventy. We can see there are some outliers.

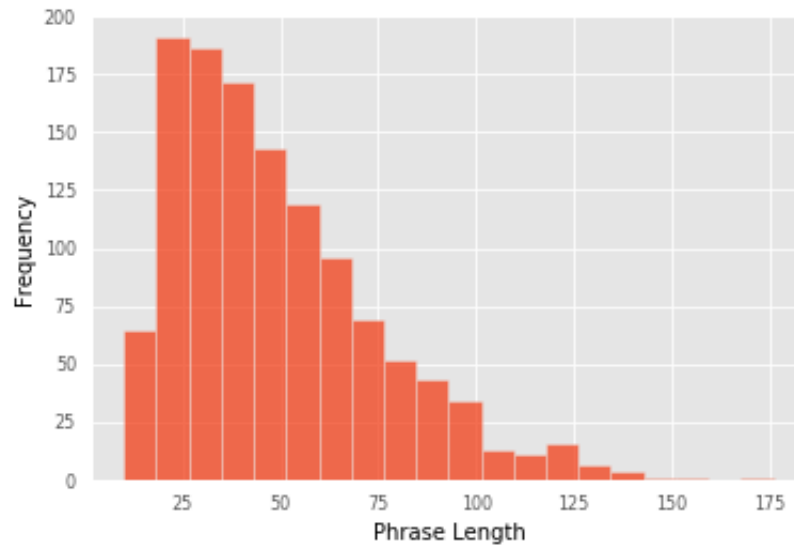


Figure 3.3.2: Phrase length of ham message

At figure 3.3.2 we can see a histogram for text length of Ham message. Ham texts are of different length. But for our dataset we can see that most of the message have text length of from 20 to 45. Again, some messages have text length more than 120. There are some short text having length of 10. As in our day to day to life most of the messages are of Ham. So, it is obvious to see variation in their text length.

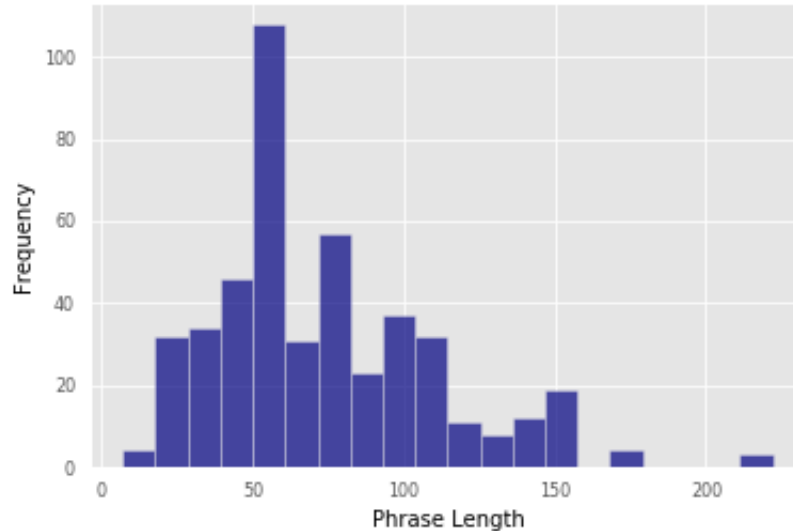


Figure 3.3.3: Phrase length of spam message

Figure 3.3.3 is histogram of spam message. Our approach is to detect spam message. So, analyzing spam data is very important. Most of the spam message are short and their length varies between 30 to 100. In our dataset we can see that most of the spam message have length between 55 to 60. We can see that there is only one spam message which have a length of more than 200.

After comparing the histogram of spam data and ham data, we found that the text length of spam is very short. Because in the most of the spam message the content is almost same. In case of ham text, we can see that it has more variation comparing to spam message.

3.4 Proposed Methodology

Our paper aims to detect spam text in Bangla language. We have taken a machine learning approach for the task. Now we are going to describe our work procedure. Our approach is mainly divided into seven sections.

- Data Pre-processing
- Data Annotation
- Data Analysis
- Feature Extraction
- Implementation of Machine Learning

Data preprocessing means processing the data according to need. We have collected data from Facebook. Without pre-processing the data will not be able to perform well. The raw data contains anomalies. And for our approach the data need to be accurate.



Figure 3.4.1: Visualization of data without preprocessing

In Figure 3.4.1 we can see the data without preprocessing. As we have collected the data from Facebook, there are different types of anomalies in the data. Using preprocessing we removed the anomalies. In Facebook, people use different types of emoji. In our machine learning-based classification we cannot work with emoji. So manually we have removed emoji from different comments. Then people perform different types of spelling mistakes on Facebook. We tried to correct the spelling. While working with data negation handling is an important task. So, we have worked with negation. Performing these operations makes our data prepare for the next steps. After completing these steps, the pre-processing of our data completes.

Data Annotation was the key part of our work. We were focused to annotate our data correctly. As our work finds out a message is spam or not so we made two categories. One category is Ham message and another is Spam message. Then we labeled the data with tags. If the message had Spam content, we labelled that as Spam and if the comment is appropriate, we labelled that as Ham. We worked in groups to annotate the data. To be fair with the data the annotation was done in three steps. At the first step, the data was labelled with one group. Then the authenticity of the label was checked by another group. Finally, the ultimate labelling was done with the collaboration of two groups. We worked that way so that we do not commit any mistake on labelling data. For each message the answer whether the message is Spam or not.

1. আজকে শনিবার

English: Today is Saturday.

This text does not contain any spam element so we have labelled it as ham.

2. আপনি লটারি জিতেছেন

English: You have won lottery.

This one is a spam message. It talks about lottery. We have labelled it as spam.

3. ইউরোপের ভিসা ৩০ দিনে

English: Get visa of Europe in just 30 days.

This sentence is a hoax. So, this one has been labelled as spam.

4. ঘরে বসেই মাসে ৩০ হাজার টাকা আয় করুন

English: Earn 30 thousand takas in a month by working from home.

This one is also a spam message.

5. দুই ঘণ্টা যাবত বৃষ্টি হচ্ছে

English: It has been raining for two hours.

This one is a ham text

6. গরিবদের সাহায্য করা উচিত

English: We should help the poor.

This one is a ham text.

Feature Extraction is very important part of our work. We have extracted the feature with count vectorizer and Term frequency-inverse document frequency vectorizer. The count vectorizer tokenizes the text and creates a vocabulary of known words. Count vectorizer encodes a new document using that vocabulary.

$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right) \quad (1)$$

Here,

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

The Term frequency-inverse document frequency is the number of times a word appears in a document divided by the total number of words in that document and the second term is Inverse document frequency, computed as the logarithmic of the number of the documents in the corpus divided by the number of documents where the specific term appears.

Implementation of Machine learning is the final step of our work. After feature extraction the key part is to applying machine learning algorithm. Machine learning model learns from the features. And we have extracted features from the data. We have implemented Machine Learning to perform our task. Supervised learning has been used. Supervised learning means supervising the data. We will train the data by using labels. By using that labels data will learn which important features make the data ham or spam. Then the trained model will predict new data. As we have collected data from Facebook and other platforms and then we have labelled the data according to their criteria. After that, we have trained our model with the labelled data. Our model has learnt from the labelled data which are spam, and which is ham. For classification we have used,

- Naïve Bayes
- Support Vector Machine
- Logistic Regression
- Decision Tree

All of them are very popular algorithm for classification problem. We divided our data into a training set and testing set. Then we fed the data to algorithms. After that, we got accuracy, precision, recall.

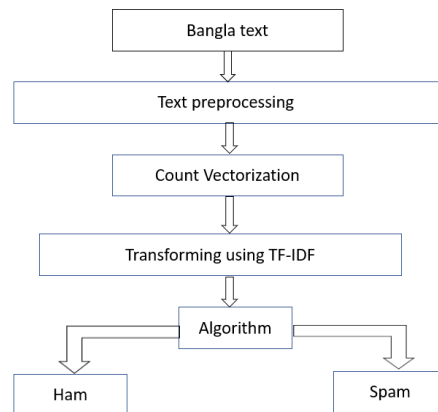


Figure 3.4.2 Flow diagram of proposed approach

The Figure 3.4.2 shows the flow diagram of our approach. Our data is Bangla text. We have preprocessed the data. Then we have used count vectorization to vectorize the data. Then we have used term frequency inverse document frequency to give emphasis to necessary feature. The we have used different machine learning algorithms to make the model.

3.5 Implementation Requirements

- Language: Python (Version: 3.7.4)
- IDE: Jupyter Notebook
- Library: Pandas (Data analysis)
- Library: Matplotlib (Data visualization)
- Library: Scikit learn (Machine learning)
- Microsoft Office

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

CPU: AMD Ryzen R7 2700x

GPU: Sapphire Radeon Rx 560 4GB DDR5

Motherboard: ASROCK X470 MASTER SLI AMD CHIPSET

RAM 8*2 3200 MHZ

Storage: 240 GB M.2

PSU: 550W 80 Plus bronze certified semi modular.

4.2 Experimental Results & Analysis

Four supervised algorithms have been used in this study. All the algorithms have their own pros and cons. During result analysis we can understand the efficiency of the algorithm on the particular task. Using different categories for performance analysis we can understand the efficiency of the algorithm. In this study to evaluate performance of the algorithms we have used:

- Accuracy
- Precision
- Recall
- F-1 score
- AUC-ROC curve

Accuracy is the summation of the true positive and true negative divided by the summation of true positive, true negative, false positive and false negative.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2)$$

Here,

TP = When the algorithm predicts positive and the result is positive.

FP = When the algorithm predicts positive and the result is negative.

TN = When the algorithms predicts negative and the result is negative.

FN = When the algorithm predicts negative and the result is positive.

Precision is the value of true positive divided by the summation of the true positive and false positive.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

Recall is true positive divided by the summation of the true positive and false positive.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

AUC-ROC curve is very important for performance evaluation of classifiers. AUC stand for area under the curve and ROC stands receiver operating characteristics. AUC-ROC curve is a performance measurement for classification problem at various thresholds setting. It represents how classifier can distinguish between classes. Higher area under the curve means the model is good.

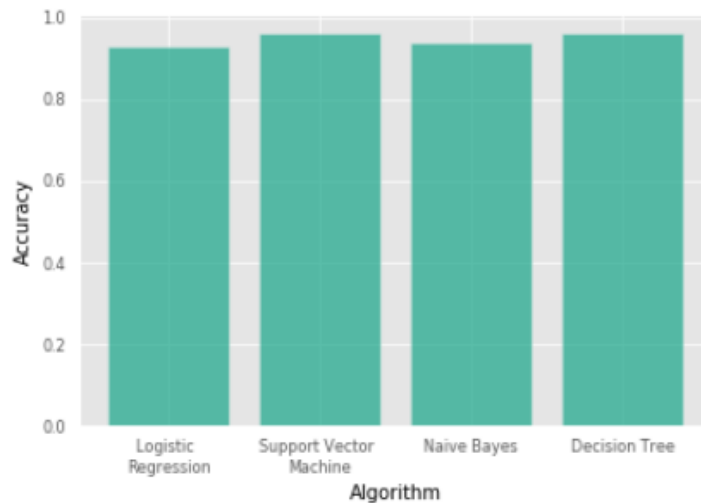


Figure 4.2.1 Accuracy of algorithms

In Figure 4.2.1 we can see the accuracy of algorithms. Logistic regression got an accuracy of 93%. Support vector machine got an accuracy of 96%. Naïve Bayes got an accuracy of 94%. Decision tree got an accuracy of 96%. We can see that the accuracy of all the algorithms are very high. But we know that accuracy can be confusing sometimes due to testing data. For example, our testing dataset only contains ham data. And the classifier is biased towards ham data or every time it just predicts ham data. So, when we will test the

classifier with just ham data. It will score a perfect accuracy. But in reality, the classifier is very poor. So, we cannot rely only on the accuracy to evaluate the algorithms. So, we have used other methodologies to evaluate the performance.

We have used four machine learning algorithms for our approach. They are

- Support vector machine
- Naïve Bayes
- Logistic regression
- Decision tree

Support Vector Machine is very popular for classification problem. As our work is based on supervised learning. We have chosen SVM as one of our algorithms. SVM basically works on hyperplane. It performs very good on small dataset.

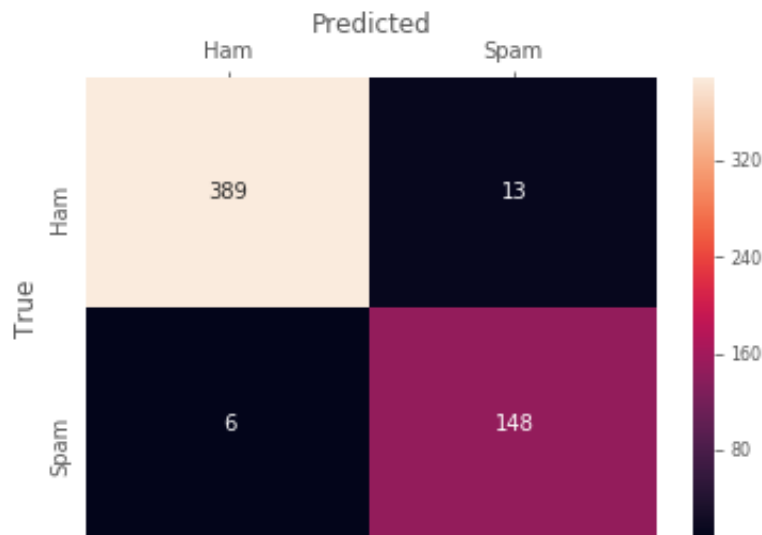


Figure 4.2.2: Confusion matrix of svm

The Figure 4.2.1 is the confusion matrix of SVM. This confusion matrix will give us a good overview of the performance of SVM. The observations are:

- SVM has correctly classified Ham for 389 out of 402 data.
- SVM has correctly classified Spam for 148 out of 154 data.
- SVM has wrongly classified 13 Ham as Spam.
- SVM has wrongly classified 6 Spam as Ham.

Table 4.2.1: classification report of support vector machine

	Precision	Recall	F1-score
Ham	0.98	0.97	0.98
Spam	0.92	0.96	0.94

Table 4.2.1 shows the classification report of support vector machine. For ham, precision is 98% and recall is 97% and f1-score is 98% which is very good. On the other hand, for spam, precision is 92%, recall is 96% and f1-score is 94%. We know recall is measurement of true positive rate. High recall means the algorithm is predicting very well. We can see in case of ham the recall is 99%. That means the algorithm predicts 99% Ham as Ham and for spam recall is 96%.

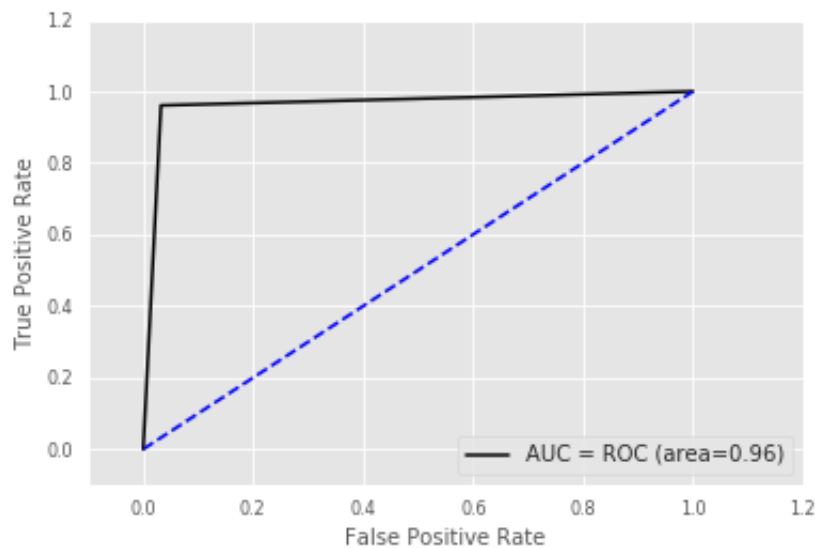


Figure 4.2.3: AUC-ROC curve of SVM

The AUC-ROC curve illustrates how well the classifier distinguish between classes. The higher area under curve the better performance of the algorithms. The Figure 4.2.3 states that the area under curve for support vector machine is 96%. So, the model can very accurately distinguish between spam and ham.

Naïve Bayes is a very popular algorithm for classification problems. It is based on Bayes theorem with an assumption of independence among predictors. Bayes theorem provides a way of calculating posterior probability. Naïve Bayes model is easy to build and particularly useful for very large datasets. It is easy and fast to predict class of test dataset.

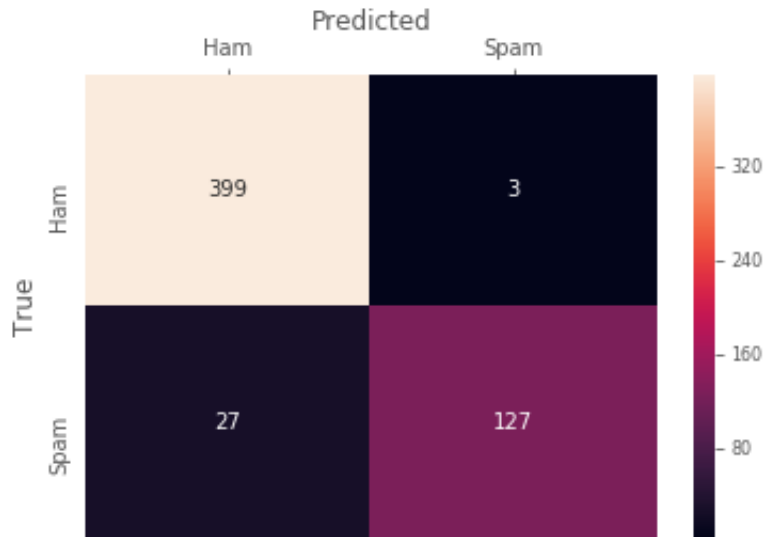


Figure 4.2.4: Confusion matrix of naïve bayes

- Naïve Bayes has correctly classified Ham for 399 out of 402 data.
- Naïve Bayes has correctly classified Spam for 127 out of 154 data.
- Naïve Bayes has wrongly classified 3 Ham as Spam.
- Naïve Bayes has wrongly classified 27 Spam as Ham.

Table 4.2.2: Classification report of naïve bayes

	Precision	Recall	F1-score
Ham	0.94	0.99	0.96
Spam	0.98	0.82	0.89

Table 4.2.2 shows the classification report of naïve bayes. For ham, precision is 94% and recall is 99% and f1-score is 96% which is very good. On the other hand, for spam, precision is 98%, that means the algorithm have identified spam data correctly during identification. But we can see that the recall is slightly low. The recall is 82%. This recall

represents that the algorithm made some mistake to identify some spam on total number of data. It has misidentified 27 spam as ham.

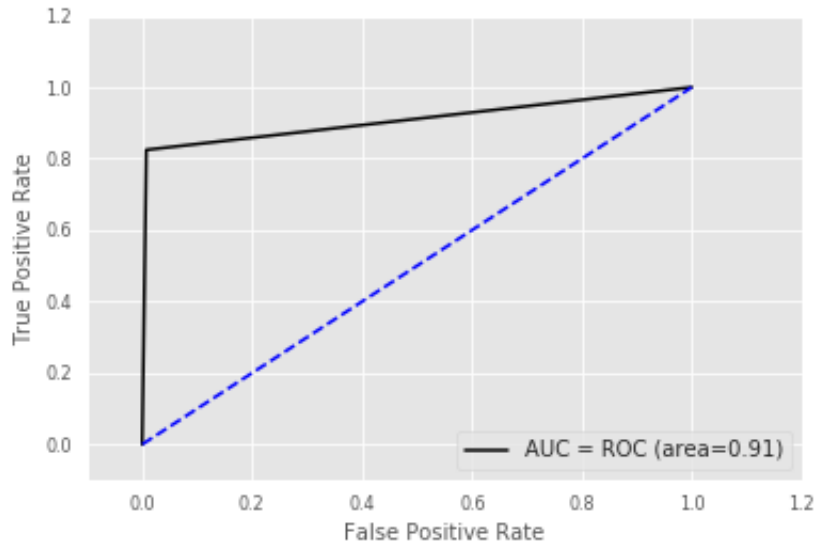


Figure 4.2.5 AUC-ROC curve of naïve bayes

Naïve Bayes classifier has good precision and recall for both ham and spam message. Good precision and recall convey how well the model has identified ham and spam data. The Figure 4.2.5 illustrates the AUC-ROC curve of the Naïve Bayes classifier. The x-axis represents false positive rate and the y-axis presents the true positive rate. As the model can identify both ham and spam very accurately the area under the curve is 91%.

Logistic Regression is a supervised machine learning algorithm. It is based on regression task. Regression finds the relation between dependent and independent variables. It predicts the probability of a target variable. As we have two class ham and spam. We have used binary logistic regression. In binary logistic regression the dependent variable is binary in nature.

Theoretically a logistic regression model predicts: $p(y=1)$

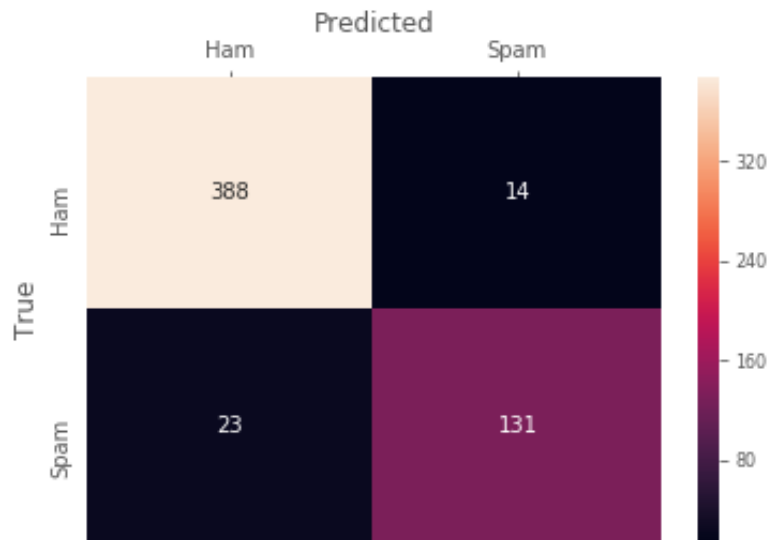


Figure 4.2.6 Confusion matrix of logistic regression

- Logistic regression has correctly classified Ham for 388 out of 402 data.
- Logistic regression has correctly classified Spam for 131 out of 154 data.
- Logistic regression has not wrongly classified 14 Ham as Spam
- Logistic regression has wrongly classified 23 Spam as Ham.

Table 4.2.3: classification report of logistic regression

	Precision	Recall	F1-score
Ham	0.94	0.97	0.95
Spam	0.90	0.85	0.88

The precision of ham is 94%, that means the percentage of positive identification of the algorithm is very good. The recall of ham is 97%, that means for ham logistic regression has identified almost all ham data.

The precision of spam is 90% which very good. The algorithm has identified good number of spam data. The recall is 85% which is very good. Recall represents what number of actual positive was identified correctly. As the percentage is very low, it represents the algorithm is unable to identify spam for the dataset. For this reason, the F1-score is very low.

Though logistic regression performs very well for ham data but the performance is low for spam data.

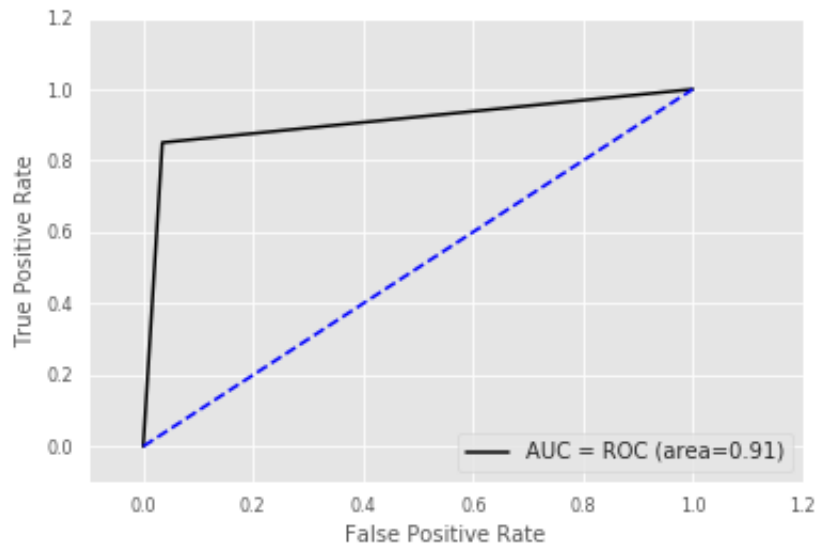


Figure 4.2.7: AUC-ROC of logistic regression

The AUC-ROC curve recounts the characterization of the classifier between the classes. The higher area under curve the better performance of the algorithms. The Figure 4.2.7 states that the area under curve for logistic regression which is 96%. So, the model can very accurately identify between spam and ham.

Decision Tree is one of the most predictive modelling approaches used in machine learning. It is based on supervised learning method.

Advantage of Decision Tree:

- Simple to use and understand.
- It can handle categorical and numerical value.
- It is resistant to outliers

Disadvantage of decision Tre:

- It is very prone to overfitting.
- Parameter tuning is complex.
- Some class can dominate due to biased learning.

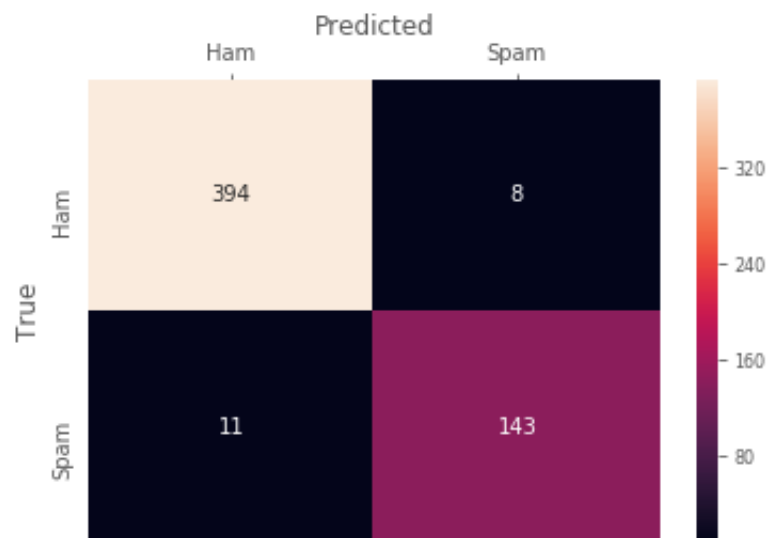


Figure 4.2.8 Confusion matrix of decision tree

- Decision Tree has correctly classified Ham for 491 out of 491 data.
- Decision Tree has correctly classified Spam for 10 out of 64 data.
- Decision Tree not wrongly classified any Ham.
- Decision Tree has wrongly classified 54 Spam as Ham.

Table 4.2.4: Classification report of decision tree

	Precision	Recall	F1-score
Ham	0.97	0.97	0.97
Spam	0.92	0.93	0.93

The precision of ham is 97%. The precision is very high. This precision shows during identification of ham decision tree has performed very well. The recall for ham is 98%. It represents that the identification of ham in all data is very high. That means the algorithm is trained well to distinguish ham from spam. As, the precision and recall are very high for ham and we know that f1-score is the average of precision and recall that is why f1- score is very high like precision and recall.

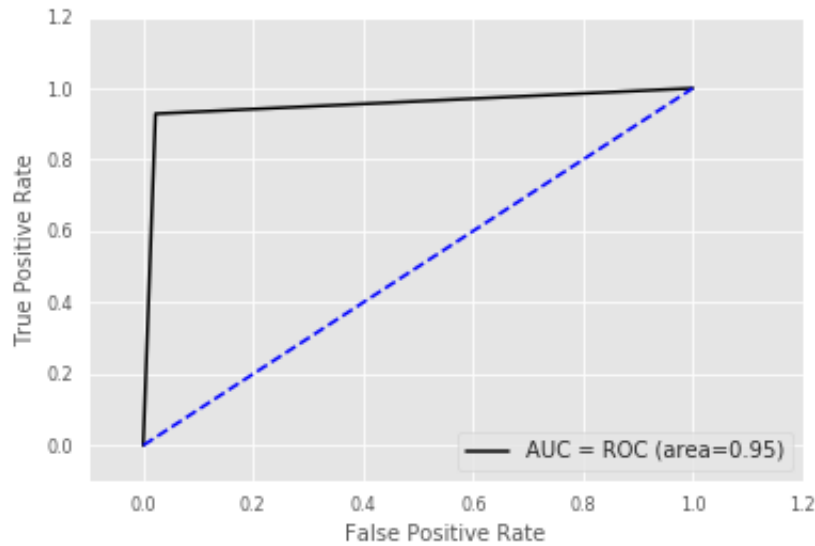


Figure 4.2.9 AUC-ROC Curve of Decision Tree

The decision tree classifier got very precise result for both ham and spam. A model of higher precision and recall can classify data very accurately. The Figure 4.2.9 illustrates the AUC-ROC curve of the decision tree classifier. The x-axis represents false positive rate and the y-axis presents the true positive rate. As the model can identify both ham and spam very precisely the area under the curve is 96%.

4.3 Discussion

In the above section we have analyzed the performance of algorithms by using accuracy, precision, recall, f1-score. Then we have shown receiver operating characteristic. Accuracy is not enough to evaluate the performance of an algorithm. Precision and recall give better understanding of algorithms performance. So, we have analyzed the precision, recall and f1-score for every algorithm.

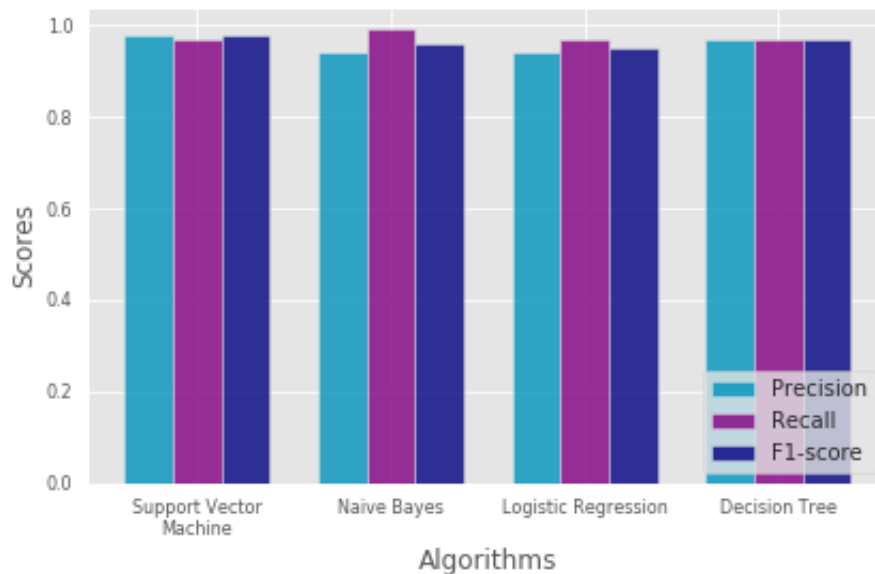


Fig 4.3.1: Performance of algorithms on ham data

The Figure 4.3.1 shows comparative analysis between the four algorithms for ham data. Support vector machine got precision of 98%. Naïve Bayes got precision of 94%. Logistic regression got precision of 94%. Decision tree got precision of 97%. We know there is a trade off relation between precision and recall. We can see that Support vector machine got recall of 97%. Naïve Bayes got recall of 99%. Logistic regression got recall of 97%. Decision tree got recall of 97%. We can see that for ham data algorithm have got good precision, recall and f1-score. But as spam classification is the main part of work. So, spam classification should be also good.

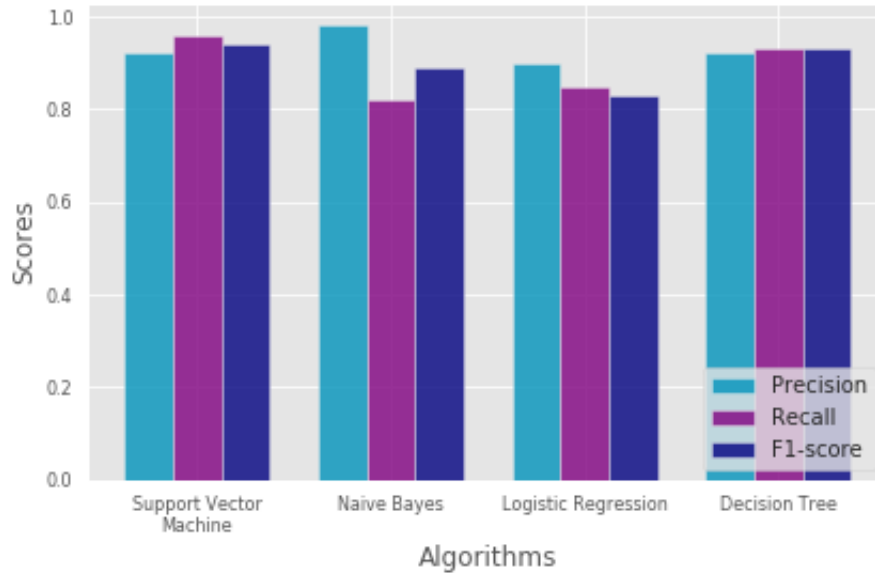


Figure 4.3.2: Performance of algorithms on spam data

The Figure 4.3.2 shows comparative analysis between the four algorithms for spam data. Support vector machine got precision of 92%. Naïve Bayes got precision of 98%. Logistic regression got precision of 90%. Decision tree got precision of 92%. On the other hand, support vector machine got recall of 96%. Naïve Bayes got recall of 82%. Logistic regression got recall of 85%. Decision tree got recall of 93%.

We can see that Naïve Bayes and logistic regression has low recall compared to precision. Precision is the amount of positive identification of the algorithm and recall is the rate of actual positive identification among total positive.

Support vector machine got f1 score of 98% on ham data and f1-score of 94% on spam data. So, in our study support vector machine gave optimal performance.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Our approach will bear great impact on society. There are many incidents of spamming in Bangla language. By using our approach, people can save themselves from Bangla spam messages. This study can reduce spam message to a great extent. Every year during board examination spam text reach at its peak in Bangladesh. Many Facebook page and groups advertise to sell question paper. In our wok, we have worked with six types of spam messages. Our work can reduce spam in Bangla language to a great extent.

5.2 Impact on Environment

Our approach can make impact on web environment. It can reduce spam message in Bangla language. As the internet user of Bangladesh is very high. Most of them uses internet in Bangla language, our can protect them for spam content. There are many victims of spam. Most of them have lost large amount of money due to spam. So, our work can secure internet user from spam content and messages.

5.3 Ethical Aspects

Our work about spam text detection in Bangla language is completely new. There is no work on Bangla spam text detection. There is no dataset available for Bangla spam text. We have made our dataset to perform our task. We have given references of all resources. Our work is ethical.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the study

In this work, we have made a dataset of Bangla spam text. There is no available dataset for Bangla spam message. Along that there is no significant work on Bangla spam text. So, we made a dataset and used machine learning to detect Bangla spam text. We have shown performance comparison between Support Vector Machine, Naïve Bayes, Decision tree and Logistic Regression. To show the comparison we have used accuracy, precision, recall and f-1 score.

6.2 Conclusions

In our work, we made a new dataset in the Bangla language for Spam text detection. We divided the dataset into two groups and labeled them. There were anomalies in the dataset, and we have processed them to remove the anomalies. After that, we have extracted features from our dataset to use in our model. Then we applied machine learning algorithms and compared their performances.

6.3 Implication for Further Study

Our approach is dependent on text. But now a days there are images that contains advertisement of spam. So, our approach will not work on image. In future deep learning can be used to detect Bangla spam text from image.

References

- [1] Dhaka Tribune, available at << <https://www.dhakatribune.com/world/2020/02/17/bengali-ranked-at-7th-among-100-most-spoken-languages-worldwide> >>, last accessed on 06-06-2020 at 7.21 PM
- [2] Wikipedia, available at <<https://en.wikipedia.org/wiki/Bengali_language>> last accessed on 12-03-2020 at 3.21 AM
- [3] BBC, available at << <https://www.bbc.com/bengali/news-47720847>>> last accessed on 28-05-2020 at 6.06 PM
- [4] WorldBank, available at <<<https://data.worldbank.org/indicator/IT.NET.USER.ZS?contextual=default&nd=2017&locations=BD&start=1990&view=chart>>>, last accessed on 29-05-2020 at 2.20 AM
- [5] Social-Media stats, available at << <https://gs.statcounter.com/social-media-stats/all/bangladesh> >>, last accessed on 21-03-2020 at 9.31 PM
- [6] Anuj Kumar Singh, Shashi Bhushan and Sonakshi Vij, “Filtering spam messages and mails using fuzzy c means algorithm,” International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU) International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), 2019.
- [7] Alexey S Katasev, Lilia Yu. Emaletdinova and Dina V. Dataseva, “Neural Network Spam Filtering Technology,” International Conference on Industrial Engineering, Applications and Manufacturing, 2018.
- [8] Alexey S Katasev, Lilia Yu. Emaletdinova and Dina V. Dataseva, “Neural Network Spam Filtering Technology,” International Conference on Industrial Engineering, Applications and Manufacturing, 2018.
- [9] Manmohan Singh, Rajendra Pamula and Shudhanshu kumar shekha, “Email Spam Classification by Support Vector Machine,” International Conference on Advanced Computer Science and Information Systems, pp.878 – 882, 2018.
- [10] Ms T. Indhumathi, R. Harshini, S. Janani and S. Navaneetha, “An Efficient Scheme for Identifying Spam Bots and Terminate Mailing,” International Conference on Science Technology Engineering and Management, pp. 34-37, 2016.
- [11] Pingchuan Liu and Teng-Sheng Moh, “Content Based Spam E-mail Filtering,” International Conference on Collaboration Technologies and Systems, pp. 218-224.
- [12] Prajakta Patil, Rashmi Rane and Madhuri Bhalekar, “Detecting Spam and Phishing Mails Using SVM and Obfuscation URL Detection Algorithm,” International Conference on Inventive Systems and Control, pp. 1-4, 2017.
- [13] Mohammad Reza Parsaei and Mohammad Salehi, “E-Mail Spam Detection Based on Part of Speech Tagging,” International Conference on Knowledge-Based Engineering and Innovation, pp. 1010-1013, 2015.
- [14] Simranjit Kaur Tuteja and Nagaraju Bogiri, “Email Spam Filtering using BPNN Classification Algorithm,” International Conference on Automatic Control and Dynamic Optimization Techniques, pp. 915-919, 2016.

[15] Harpreet Kaur and Ajay Sharma, “Improved Email Spam Classification Method Using Integrated Particle Swarm Optimization and Decision Tree,” International Conference on Next Generation Computing Technologies, pp. 516 – 521, 2016.

[16] BTRC, available at << <http://www.btrc.gov.bd/content/internet-subscribers-bangladesh-february-2020>>>, last accessed 05-04-2020, 7.21 PM.

PLAGIARISM REPORT

Submission date: 06-Jul-2020 06:12PM (UTC+0600)

Submission ID: 1354099091

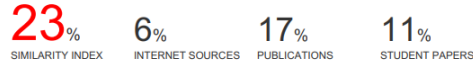
File name: plagiarism_Bangla_spam_text_detection.pdf (584.54K)

Word count: 7282

Character count: 35297

Bangla spam text detection

ORIGINALITY REPORT



PRIMARY SOURCES

1	Shovon Ahammed, Mostafizur Rahman, Mahedi Hasan Niloy, S. M. Mazharul Hoque Chowdhury. "Implementation of Machine Learning to Detect Hate Speech in Bangla Language", 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), 2019 Publication	12%	5	Submitted to University of Maryland, University College Student Paper	1%
2	Submitted to Daffodil International University Student Paper	2%	6	tessera.spandidos-publications.com Internet Source	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%	7	Akhtar, Li, Pei, Imran, Rajput, Azeem, Wang. "Diagnosis and Prediction of Large-For-Gestational-Age Fetus Using the Stacked GeneralizationMethod", Applied Sciences, 2019 Publication	<1%
4	Pankaj R. Chandre, Parikshit N. Mahalle, Gitanjali R. Shinde. "Chapter 5 Deep Learning and Machine Learning Techniques for Intrusion Detection and Prevention in Wireless Sensor Networks: Comparative Study and Performance Analysis", Springer Science and Business Media LLC, 2020 Publication	1%	8	Submitted to CVC Nigeria Consortium Student Paper	<1%
	Classification Techniques", International Journal of Information Technology and Computer Science, 2016 Publication		9	Submitted to University of Portsmouth Student Paper	<1%
15	Submitted to University of Nottingham Student Paper	<1%	10	Submitted to CSU, San Jose State University Student Paper	<1%
16	link.springer.com Internet Source	<1%	11	Submitted to University of Sheffield Student Paper	<1%
17	Submitted to Napier University Student Paper	<1%	12	Submitted to Troy University Student Paper	<1%
18	"InECCE2019", Springer Science and Business Media LLC, 2020 Publication	<1%	13	Submitted to King's College Student Paper	<1%
19	hdl.handle.net Internet Source	<1%	14	Siddu P. Algur, Prashant Bhat. "Web Video Mining: Metadata Predictive Analysis using 2017 Publication	<1%
20	"ICDSMLA 2019", Springer Science and Business Media LLC, 2020 Publication	<1%	23	uksim.info Internet Source	<1%
21	Submitted to University of Pretoria Student Paper	<1%	24	toc.proceedings.com Internet Source	<1%
22	Argyro Mavrogiorgou, Athanasios Kiourtis, Dimosthenis Kyriazis. "Chapter 6 A Comparative Study of Classification Techniques for Managing IoT Devices of Common Specifications", Springer Science and Business Media LLC, Publication	<1%	25	ami.info.umfcluj.ro Internet Source	<1%
			26	inccst.muet.edu.pk Internet Source	<1%
			27	Submitted to The Robert Gordon University Student Paper	<1%
			28	Submitted to City University of Hong Kong Student Paper	<1%
			29	worldcomp-proceedings.com Internet Source	<1%
			30	Dilip Kumar Choubey, Sanchita Paul. "chapter 12 GA_SVM", IGI Global, 2017 Publication	<1%
			31	Submitted to University of Northumbria at Newcastle Student Paper	<1%
			32	Submitted to University of Melbourne Student Paper	<1%

33	researchonline.lshtm.ac.uk Internet Source	<1%
34	Submitted to The Hong Kong Polytechnic University Student Paper	<1%
35	www.slideshare.net Internet Source	<1%
36	"The International Conference on Advanced Machine Learning Technologies and Applications (AMLT2018)", Springer Science and Business Media LLC, 2018 Publication	<1%
37	Submitted to Institute of Research & Postgraduate Studies, Universiti Kuala Lumpur Student Paper	<1%
38	Submitted to ABV-Indian Institute of Information Technology and Management Gwalior Student Paper	<1%
39	"Intelligent Computing, Networking, and Informatics", Springer Science and Business Media LLC, 2014 Publication	<1%
40	"Cognitive Informatics and Soft Computing", Springer Science and Business Media LLC, 2019 Publication	<1%

Exclude quotes Off
Exclude bibliography Off
Exclude matches Off