

**USING SOCIAL NETWORKS TO DETECT
MALICIOUS BANGLA TEXT CONTENT**

BY

**TANVIRUL ISLAM
ID: 152-15-6117**

**NAZMUL ISLAM
ID: 152-15-5652**

**SK. MEHEDI HASSAN
ID: 161-15-7667**

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

Supervised By

Ms Subhenur Latif
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

APRIL 2019

APPROVAL


This research titled “Using Social Networks to Detect Malicious Bangla Text Content”, submitted by Tanvirul Islam, ID No: 152-15-6117, Nazmul Islam, ID No: 152-15-5652, SK. Mehedi Hassan, ID No: 161-15-7667 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 3rd May, 2019.

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. Md. Ismail Jabiullah
Professor

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

Internal Examiner


Dr. Sheak Rashed Haider Noori
Associate Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner


Dr. Dewan Md. Farid
Associate Professor
Department of Computer Science and Engineering
United International University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Ms Subhenur Latif, Assistant 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:



Ms Subhenur Latif
Assistant Professor
Department of CSE
Daffodil International University

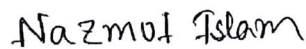
Co- Supervised by:

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate
Head
Department of CSE
Daffodil International University

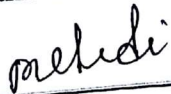
Submitted by:



Tanvirul Islam
ID: 152-15-6117
Department of CSE
Daffodil International University



Nazmul Islam
ID: 152-15-5652
Department of CSE
Daffodil International University



SK. Mehedi Hassan
ID: 161-15-7667
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to Ms. Subhenur Latif, Assistant Professor, 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. Her 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 at Daffodil International University, who took part in this discussion while completing the course work.

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

ABSTRACT

Social spam has rapidly increased over recent years. This kind of spam contents like text messaging or comments has a massive negative influence on normal user's experience in social media. Our research reflects the current experimental study on spam detection from different textual data. In this experimental research, we have used Naïve Bayes classifier, a supervised machine learning algorithm with feature extraction to detect spam from Bangla text at the sentence level. We have started this research by collecting Bangla textual data from YouTube, Facebook, and other social media. Then we have categorized the sentences into two polarities i.e. spam and ham applied to the Multinomial Naïve Bayes classification algorithm. Our proposed system detects spam on the basis of the polarity of each sentence associated with it. Finally, our experiment shows that the model has an accuracy of 82.44% in detecting spam Bangla text content.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of Examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
List of Figures	vii
List of Tables	viii
CHAPTERS	
Chapter 1: Introduction	1-3
1.1 Introduction	1
1.2 Motivation	1
1.3 Rationale of the Study	2
1.4 Research Question	3
1.5 Expected Output	3
1.6 Report Layout	3
Chapter 2: Background	4-7
2.1 Introduction	4
2.2 Related Works	4
2.3 Research Summary	6
2.4 Scope of the Problem	6
2.5 Challenges	7
Chapter 3: Research Methodology	8-13
3.1 Introduction	8
3.2 Research Subject and Instrumentation	8
3.2.1 Research Subject	8

3.2.2 Instrument	9
3.3 Data Collection Procedure	10
3.4 Methodology and Data Analysis	11
3.4.1 Pre-Processing	11
3.4.2 Feature Extraction	12
3.4.3 Training	12
3.4.4 Algorithm	12
3.5 Implementation Requirements	13
Chapter 4: Experimental Results and Discussion	14-16
4.1 Introduction	14
4.2 Experimental Results	14
4.3 Summary	16
Chapter 5: Summary, Conclusion, Recommendation and Implication for Future Research	17-18
5.1 Summary of the Study	17
5.2 Conclusions	17
5.3 Recommendations	17
5.4 Implication for Further Research	18
5.5 Future works	18
REFERENCES	19

LIST OF FIGURES

FIGURES	PAGE
Figure 3.4.1 Overall procedures of text classification in terms of a block diagram	11
Figure 3.4.2 Naïve Bayes Algorithm (Multinomial Model): Training and Testing	13
Figure 4.2.2 Confusion Matrix in Bar Chart	15
Figure 4.2.4 ROC Curve	16

LIST OF TABLES

TABLES	PAGE
Table 3.1 Text category based on interpretation	09
Table 3.3.1 Bangla Spam Dataset	10
Table 4.2.1 Confusion Matrix	14
Table 4.2.1 Performance Measure	15

CHAPTER 1

INTRODUCTION

1.1 Introduction

Social Media plays an important role in communications in digital Bangladesh. Social network sites are basically represented by Facebook, Twitter, YouTube, and many others. At present, people spend lots of time on social media. Like Celebrities, public figures, business icons create their social pages for interacting with online users and their fans. But lots of malicious behaviors on social media makes many troubles to the users.

Social Networks (SNs) have become an important part of user's social identity. The initial intent of SNs was to facilitate the connection and sharing. So, People are heavily dependent on online interactions for communications. The increases in content in social media are responsible for the increases in social spams. Social networking platform do not offer Satisfactory validation mechanisms to identify the spammers. Professionals guess that almost 40% of SNs accounts are involved with spam [1].

So in the beginning, we collected lots of samples of spam and ham from real-life uses in social media to produce the training dataset. After that a filtering method of naive Bayes classification will be described and will be applied this method on created dataset for testing. These model will be evaluated through tasting will be discussed. Though numerous machine learning algorithms are exist but probably due to their simplicity and accuracy, we implement the Multinomial Naïve Bayes (MNB) algorithm, an accuracy rate of 82.44%.

1.2 Motivation

Social networks have lifted the communication system to the utmost level. People spend most of their times Facebook, Twitter, YouTube etc. rather than search engines. In Bangladesh and people who speak Bangla, most of them prefer using Bangla for networking and developing communication. Content sharing, content contribution, comment or other feedback system is being used to interact with online users by business entities or other celebrities. They create their social pages to enrich communication. During

the time, social media become vulnerable to diverse types of undesirable and hateful spam text contents. Spammers destroy the network environment and this degrades the user's experience of using the network. There is an important necessity in the civilization for saving the image and continue a well environment in social media. That's why we have decided to work on spam Bangla Text content. We have gone through many spam detecting research papers which had been done in English. Work on Bangla language is very few to detect emotion. So, we thought to work in Bangla spam text. In this experiment, we intend a social media based Bangla spam content detection mechanism.

1.3 The Rationale of the study

Due to textual complexity, identifying spam text from Bangla language is not easy. There have been lots of researches for analysis spam text content in English Language which demonstrated satisfactory outcomes. This become possible because tons of textual data are available on World Wide Web during this period. Prior to this time, it was not was to create a dataset of spam and train a model using this dataset. Still, spam Bangla text content analysis is a new area and have scope of improvement. Many Bangla text contains malicious links which misguide the users to fraud and phishing websites. We collected almost two thousand Bangla sentences consisting both positive and negative content. We categorized them into two polarities: spam noted by 1 and ham noted by 0. For example, “মেডাম ছাত্রের সাথে যা করল, দেখুন ভিডিও সহ”, “ভিডিওটি একা একা দেখবেন কিন্তু” this two sentences are spam links which redirect the users to a false news or websites. Like these, we have also collected ham like “ভাল হইতে পয়সা লাগেনা”, “এরাই দেশের ভবিষ্যৎ” to train the data sets. Then we applied multinomial naïve Bayes classifier to complex pattern recognition and approximation of the function.

1.4 Research Question

Question 1: Does every Bangla sentence have distinct exposition e.g. positive and negative?

Question 2: Does every negative interpretation of the sentence contain spam?

Question 3: Does spam sentence contain some specific word?

Question 4: Can we identify Spam from every new generated informal Bangla sentences?

1.5 Expected Outcome

As there have no work been done to detect malicious Bangla Text content, we have decided to go for it. Our expected output is very satisfactory. First, we will train whether the sentence is spam or ham through MNB algorithm and after the analysis, it will detect whether it is spam or ham.

1.6 Report layout

The paper is organized into five sections. Following this introduction, Chapter 2 provides brief background details of spam detection field from an information systems perspective, a survey on text analysis those have been published in different information system journals, also the scope of the problem and its challenges. A detailed description of the research methodology including the procedure of data collection, pre-processing and feature extraction is provided in chapter 3. Chapter 4 presents the experimental result of the applied methodology, a brief description of the analysis. And finally, Chapter 5 describes the summary of the empirical research, important limitations of the approach, the implication for Further Study.

CHAPTER 2

BACKGROUND

2.1 Introduction

Extensive researches have been done for the spam detection of English texts which indicated satisfactory results. Still, spam Bangla text content analysis is a new area and has scope of improvement. There are lots of Bengali native speakers, over 160 million, and massive quantities of Bangla text are produced in online social media, blogs and websites. As a result, it would be easier to check the polarity; how much positive or negative the sentence is. After analyzing the text pattern, the sentence could be categorized according to the polarity it belongs to.

In our research, we have mainly researched on how we can detect whether the sentence is spam or ham from a given Bengali text which has been collected from social media such as YouTube, Facebook, FB group like DSU, Hero Alom etc. We collected lots of samples of spam and ham from the real-life uses in social these media to produce the training dataset. After that a filtering method of naive Bayes classification will be described and will be applied this method on created dataset for testing. These model will be evaluated through testing will be discussed. Though numerous machine learning algorithms exist but probably due to their simplicity and accuracy, we implement the Multinomial Naïve Bayes (MNB) algorithm.

2.2 Related Works

The increasing popularity and the vast quantity of information exist in social network attracts researchers to study issues encountered by social media users. A large proportion of work has been done on various problems like spam detection and filtering, community detection, and information diffusion. Wuying Liu et. al. [2] proposed a novel index-based online text classification method using six individual classifiers to multiple text features of Chinese message. Houshmand Shirani-Mehr [3] applied several machine learning algorithms on an SMS Spam dataset from UCI Machine learning repository where the evaluation showed that 10-fold cross-validation showed the best accuracy in their work. A

new supervised machine learning approach has been introduced by Liu et. al.[4] to detect spam accounts in social networks through their behavioral characteristics. They collected messages crawling from Sina Weibo and then selected three categories of features extracted from message contents, social interactions and user profile properties and applied the ELM based spam accounts detection algorithm. To improve the performance of detecting spam SMS, Amir et. al. [5] proposed new content-based features. The effectiveness of the proposed features are validated using several classification methods and the results demonstrate that the proposed system can improve the performance of SMS spam detection. Moreover, X. Jin et. al.[6] proposed a scalable and online social media spam detection system for social network security. For dealing with the scalability and real-time detection challenges, they implemented GAD clustering algorithm for large scale clustering and integrated it with the designed active learning algorithm. Another approach is presented by Ashis et. al. [7] who explored the four supervised learning Methods, namely Decision Tree, KNN, Naïve Bays, and Support Vector Machine (SVM) for categorization of Bangla web documents. In their papers, Sahami et al.[8] used a Naïve Bayes classifier with a multi-variate Bernoulli model, a form of NB which relies on Boolean attributes. On the other hand, Pantel et. al.[9] adopted the Multinomial form of Naïve Bayes that normally takes into account term frequencies. It has been shown experimentally in [10] that Multinomial Naïve Bayes performs generally better than the Multivariate Bernoulli NB in text classification. An experiment has been adopted in [11] which emulates incremental training of personalized spam filters. Mohammad Samman Hossain et. al. [12] applied several classification algorithms like Logistic Regression, Boosted Tree and SVM to classify sentiment (Negative or Positive) of Bengali newspaper headline and SVM performs better having higher accuracy. In [13], a text processing approach for semantic analysis and context detection is proposed. They evaluated their approach with a public, real and non-encoded dataset along with several established machine learning methods which can enhance instant messaging and SMS spam filtering.

2.3 Research Summary

Research is an organized way to find solutions to existing problems or problems that nobody has worked on before. It can be used for solving a new problem or it can be the expansion of past work on any particular field. Our Research is on detecting spam Bengali text that is associated with NLP(Natural Language Processing).AI(Artificial Intelligence) is challenging the human being to exceed human beings performance. There's been lots of work that has already done to detect spam using texts or documents from various languages. We have studied lots of paper related to detecting spam from a text, lyrics, sentence etc. They used different methods and among them, we have chosen multinomial Naïve Bayes classification algorithm for spam text detection. For that reason, we collected lots of samples of spam and ham from real-life uses in social these media in order to produce the training dataset. After that a filtering method of naive Bayes classification will be described and will be applied this method on created dataset for testing. These model will be evaluated through tasting will be discussed. Though numerous machine learning algorithms are exist but probably due to their simplicity and accuracy, we implement the Multinomial Naïve Bayes (MNB) algorithm.

2.4 Scope of the Problem

Detecting spam from a text is incipiently a content-based classification which expatiate the concept from Natural language processing (NLP) including Machine Learning(ML) as well. The study of spam detection is very necessary. The increasing of social network, along with the faith they have in their social profile, makes a favorable situation for spammers. In fact, social spam are increasing rapidly over time. Enhanced accuracy and reliability in text mining procedures can resolve the complications. Currently, as the next wave of knowledge discovery, text analysis is achieving high commercial values. In this research, we will analyze Bengali text from Facebook status, YouTube comments etc for finding associated spam of each sentence like positive, negative. After identifying the polarity of each sentence we will then try to find spam text content of each sentence.

2.5 Challenges

Detecting spam or ham from Bangla text content provides huge challenges. some of the sentences like “তারে দেখতে তো ব্যাঙ্গের বাচ্চার মতো দেখায়, সে নাকি আবার হিরো আলম”. Here “তারে দেখতে তো ব্যাঙ্গের বাচ্চার মতো দেখায়”- is used as abuse. It indicates negativity of the sentence and the system will automatically detect as spam (1). On the other hand, “সে না কি আবার হিরো আলম” is just a simple sentence which systems detects as ham (0). Though the whole passage indicates a spam behavior, it is very complex to recognize the pattern of each word and sentences. Misspelling, stop words like ‘,’ ‘!’, ‘?’, ‘.’, ‘~’, ‘||’, ‘|’ etc degrades the processing which provides a low accuracy rate. Bangla language having a huge vocabulary, words having different meaning and their various uses makes it more complex for text mining. That provides another challenge for modeling Bangla text. Because we are doing our research based on the generated expression of the sentences, it is possible to have the same expression with different polarity.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, you will get an overview of research methods that were carried out to detect spam from a given Bengali Text. It provides information about how data can be processed by applying some certain techniques to sort out spam from them. The instrument that is used to extract the spam from Bengali text from Facebook status and other sources is also described and the procedures that were followed to carry out this data extraction are included. It also provides the methods used to analyze the textual data. Lastly, the implementation and requirements that were followed in the process are also discussed.

3.2 Research Subject and Instrumentation

3.2.1 Research Subject

The main objective of this study is to detect spam from a given Bengali text in order to come up with spam detection associated with it by using Multinomial Naïve Bayes classification algorithm. In case of finding the spam of a sentence, text mining analysis can make it very specific. A set of data is been collected and we categorized them into two sections i.e. Spam and Ham. Spam column contains Spam sentences, words whereas Ham column contains sentences which have a positive meaning. Then we used 80% of these datasets for training and else for testing.

Table 3.1 implies the categorization of spam and ham data which have been collected from various social applications' status, comments etc. It is observed that the polarity of different sentences is generated according to the categorization shown in the table 3.1 below:

Table 3.1 Text category based on interpretation

Spam (1)	Ham(0)
ভিডিওটি একা একা দেখবেন কিন্তু	
অপু বিশ্বাসের গোপন ভিডিও	
	মজা পাইছি
দৌলতদিয়ার কর্মীর গোপন ভিডিও	
	ভালই সেলিব্রেটি হইতেছেন।
গ্রামের মেয়েরা দেখুন কি করে। গোপন ভিডিও ফাস	
	শিক্ষিত নয় সুশিক্ষিত হও
	জোর যার মুল্লুক তার
ছাগলের তিন নম্বর বাচ্চা	

Whenever a Bangla Sentence is used as input, the system would possibly able to determine whether it is Spam data(1) or Ham data(0) behind the textual content based on the interpretation of sentence pattern. In this experimental study, we have presented the feature extraction technique for detecting malicious Bangla text content.

3.2.2 Instrument

For research purposes, we have collected around 2000 Bengali sentences from different sources like Facebook status, YouTube comments, textbooks, newspaper, direct speech

etc. Our work is to detect spam from a sentence by applying text classification algorithm. Some well-performed algorithm like ELM, keyword spotting method, support vector machines(SVM), hidden Markov model etc. are used in case of text analysis. Therefore these algorithms give a very high accuracy of almost 90%. In our research, we have used “Multinomial Naïve Bayes” classification algorithm to find the polarity of our test sentences.

3.3 Data collection procedure

Even though many datasets in the different language are available in the different databank for research purposes, for Bangla language no dataset are available. So, we have elect to create our own Dataset from various social media like Facebook, YouTube and named it Bangla Spam Dataset as presented in table 3.3.1.

Table 3.3.1: Bangla Spam Dataset

Total Instance	1965
Spam	1319
Ham	646

In order to come up with accurate and objective findings, A good research mainly relied on both primary and secondary data. Primary data’s are the raw data which is mainly used for the original purpose. Those data contained many stop words, punctuation and special symbols which is directly taken from the field by interviews and questionnaires. We removed those symbols and punctuation to get the secondary datasets. Secondary data is collected for purposes other than the original use. The research has been carried out using secondary data. The main intention was to create a properly trained data set consists of Bengali spam keywords.

3.4 Methodology and Data Analysis

Before applying classification algorithm, it is important to make a suitable datasets. At the same times, pre-processing of Bangla text also prerequisite previously trainings and construction of model for effective text classification. Figure 3.4.1 demonstrates the whole method of classification process.

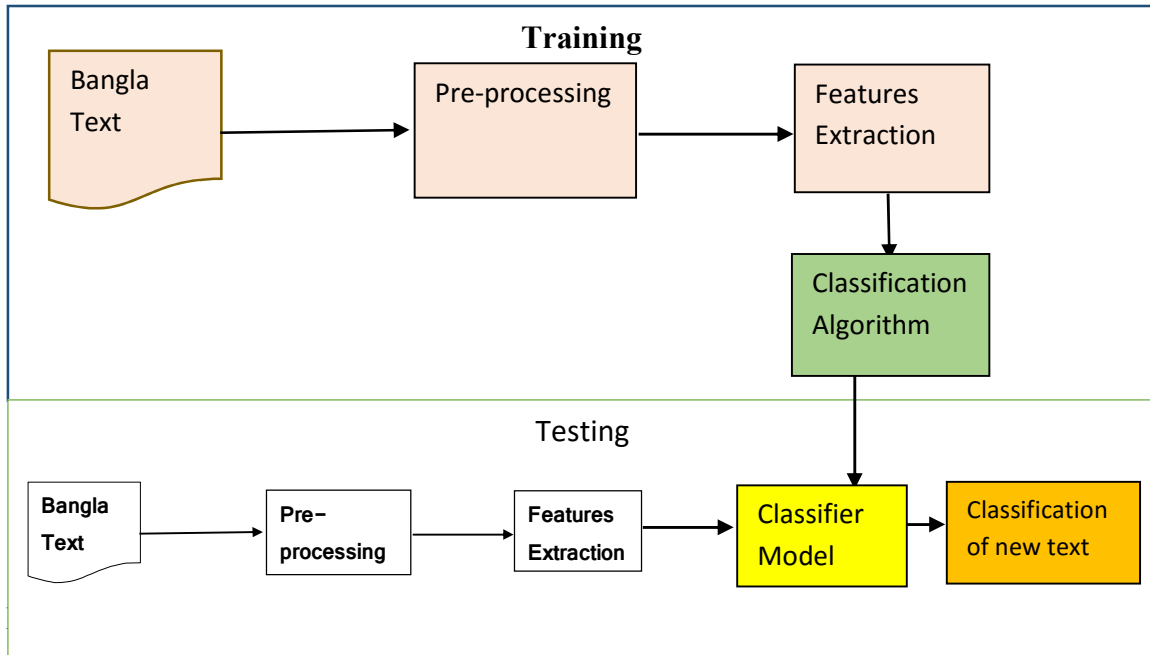


Figure 3.4.1 illustrates the detail procedures of text classification in terms of a block diagram.

3.4.1 Pre-Processing

Selecting a proper representation of data within documents is essential to obtain good Classification performance. Our dataset contains 4 column “Text”, “spam”, “related ham” & “references”. All the special symbols like (<, >, :, {, }, &, *, (,), | etc.), punctuation marks are removed from the text documents. Then we take all the spam and ham into a single column and add another column named “status” to keep the track if the text is spam or not. We keep 1 in status column for spam, and 0 for ham. Then we convert the .xlsx file in .csv file. Now our dataset is ready to load in Jupiter notebook.

3.4.2 Feature Extraction

Afterward the pre-processing, now documents have lesser amounts of words, now features extraction from this documents become easier. Several statistical approaches are being used for features extraction from text data. Here TF-IDF (term frequency–inverse document frequency) have been used to extract features from text corpus. Because it provides a way to score the importance of word based on how frequently they appear across multiple documents.

3.4.3 Training

In this section, the processed data are split into a test set and a train set. The 80/20 split is chosen. This research implements the algorithm called Naive Bayes Classifier which is used widely for spam detection. When dealing with text, it's very common to treat each unique word as a feature and since the typical person's vocabulary consist of numerous thousands of words, this makes a huge number of features. The simplicity of the algorithm and the capability of handling a large number of features makes Naïve Bayes Classifier a strong performer for classifying texts. It becomes a huge challenge regarding which variant of Naïve Bayes should be applied. It has been shown experimentally in [15] that Multinomial Naïve Bayes performs better than the Bernoulli Naïve Bayes and Gaussian naive Bayes in case of text classification. Multinomial naive Bayes treats features as event probabilities. Multinomial Naïve Bayes classifier are applied on the proposed dataset.

3.4.4 Algorithm

Naive Bayes Classifier works based on Bayesian theorem. Multinomial and Bernoulli distributions are popular while classifying document classification including Spam Filtering. In Our case, Multinomial NB do better than Bernoulli. The Multinomial NB work as follows:

```

TRAINMULTINOMIALNB(C, ID)
1  V ← EXTRACTVOCABULARY(ID)
2  N ← COUNTDOCS(ID)
3  for each c ∈ C
4  do Nc ← COUNTDOCSINCLASS(ID, c)
5     prior[c] ← Nc/N
6     textc ← CONCATENATETEXTOFALLDOCSINCLASS(ID, c)
7     for each t ∈ V
8     do Tct ← COUNTTOKENSOFTERM(textc, t)
9     for each t ∈ V
10    do condprob[t][c] ←  $\frac{T_{ct}+1}{\sum_{d'}(T_{d't}+1)}$ 
11 return V, prior, condprob

APPLYMULTINOMIALNB(C, V, prior, condprob, d)
1  W ← EXTRACTTOKENSFROMDOC(V, d)
2  for each c ∈ C
3  do score[c] ← log prior[c]
4     for each t ∈ W
5     do score[c] += log condprob[t][c]
6  return arg maxc ∈ C score[c]

```

Figure 3.4.2: Naive Bayes Algorithm (Multinomial Model): Training and testing [14].

3.5 Implementation requirement

We have used Python language for implementation where the platform is Anaconda. The tools are listed following:

1. Anaconda.
2. Python.
3. MS Excel.
4. Notepad++.
5. Socialfy. (Facebook Comment Extractor tools)
6. ytcomments. (YouTube Comment Extractor tools)

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

This is an experimental based research that we have worked out. In this chapter, the results of malicious spam text content from the Bangla language are presented according to their polarity. In total, we have collected 1965 sentences from Facebook statuses, YouTube comments, Bengali blogs, newspapers, and textbooks. The experiment has been carried by using Naïve Bayes (NB) which includes Pre-processing, feature extraction and finally text classification methods. According to their pattern polarity, spam, and ham, the sentences were identified and the results have been discussed which includes the total accuracy of our experiment in details. After evaluating the polarity results we have finally come out with a satisfactory outcome.

4.2 Experimental Results

After training, using training dataset, it's time to test our dataset using test set that is unknown to our model. From results, Multinomial Naive Bayes yields an overall accuracy of 82.44%. The confusion matrix is shown as follows in the table 4.2.1.

Table 4.2.1: Confusion Matrix

	Predicted Yes	Predicted No
Actual Yes	269	11
Actual No	58	55

The Confusion Matrix are illustrate using bar chart in following figure

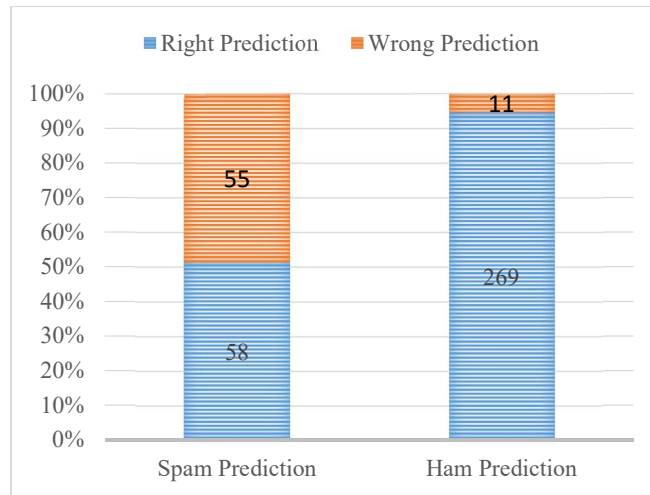


Figure 4.2.2 Confusion Matrix in Bar Chart

Error, Precision, Recall, F1 Score, AUC (Area Under Curve) are shown in the following table.

Table 4.2.3: Performance Measure

Precision	0.825
Recall	0.824
F-Score	0.808
Error	17.56
AUC (Area Under Curve)	0.72

The ROC is shown in the following figure 4.2.4

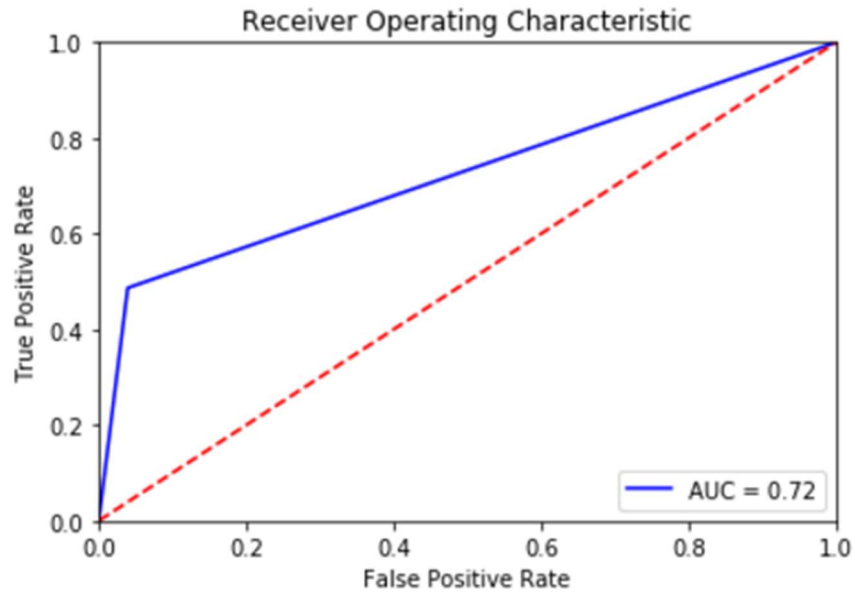


Figure 4.2.4 The ROC (Receiver Operating Characteristics) curve.

4.3 Summary

During the implementation of our system, we noticed that the bigger the number of sentences, the higher are the recall and precision. Therefore, we believe that the enrichment of our database of Bangla sentences can significantly enhance the results. After experimenting we have found that a sentence may have spam, or it may be ham. Featuring the extraction and text classification, we have used the “Multinomial Naïve Bayes” Algorithm and after the experimental result, we have come up with 82.44% accuracy.

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the study

During the last few decades' text classification has received an incredible attention from people because it helps to classify spam data and threats. Hence, a lot of work is being done in this domain to find the finest classifier for text classification. From the acquired results in contrast with the pre-processing technique, it is clear that the framework with Multinomial Naive Bayes algorithm execute better as compared to other classifiers. After extensive pre-processing MNB was applied and it comes out to be 82.44 % effective in classifying malicious Bangla Text content.

5.2 Conclusion

Detecting spam from a Bengali sentence was not that easy as people disagree on identifying exact interpretation of the same sentence. Our text classification method helps us to detect exact expression that majority people think about. Among different approaches, we have used multinomial naïve Bayes classification algorithm to extract semantic information from a sentence for detecting spam from Bangla text content. Finally, the accuracy came 82.44%.

5.3 Recommendation

In this thesis, we have worked with around two thousand sentences. So, our corpus doesn't have sufficient lexicons. As every day, new data are generating through social media, a collection of new pattern sentences are necessary. So before going for test add necessary keywords to the database. While giving input keep a focus on the spelling of the lexicons and also the removal of digit, punctuation, special symbols and stemming is very important to get the best accuracy. In the case of a spelling mistake, the program will fail to detect spam accurately. So the user may get lower accuracy.

5.4 Implications for Further Research

The demand for data mining analyst is highly appreciated in this modern age. This is because of the presence of abundant amount of data in our surroundings. To be more accurate, it is high time to work with these sorts of complex data, so that a new pattern can be introduced to resolve several critical problems. Spam analysis is one of the fundamental branches of data mining. The experimental study which we have carried out on malicious text detection with a satisfactory outcome is leaving a strong footprint behind our work. It has been observed that works on spam detection in Bangladesh a lot of valuable impact in our day to day life. We are living in the 3rd world's modern age. In this modernized world, people are seen very active in social media like Facebook, YouTube etc. Business entities set up their public pages on social networks and enhance their direct interaction with their customers through content sharing, commenting, or through any other feedback system. Celebrities, online sellers or institutional organization also publish their content for direct interaction. It is unfortunate that some spammers spoil the environment by posting unethical staffs or posting abusive comments in their posts which destroys the images. So it is very urgent to stay safe from malicious trap. Prevention of this kind of spam needs to be executed as soon as possible. As there has been no work done for Bangla languages, so our research will bring a revolutionary changes in the field of data science and to Bangladeshi people's perspective. We will further research for detecting and fighting against the spam accounts through this process.

5.5 Future Work

- ✓ Achieving higher accuracy by using classifiers in combination
- ✓ Developing a technique that can catch the sentimental phrases and train methodology for those spams.
- ✓ Multilingual spam email classification
- ✓ Enriching corpus with more words.
- ✓ Add a stemmer to reduce the size of our corpus and improve model performance.
- ✓ Detecting and fighting spam accounts.

REFERENCES

- [1] Go.proofpoint.com, 2018. [Online]. Available: <https://go.proofpoint.com/nexgate-social-media-spam-research-report>. [Accessed: 10- Sep- 2018]
- [2] W. Liu and T. Wang, "Index-based Online Text Classification for SMS Spam Filtering," *Journal of Computers*, vol. 5, no. 6, 2010.
- [3] Houshmand Shirani-Mehr, "SMS spam detection using machine learning approach", pp. 1-4, 2013.
- [4] Liu, C., Wang, G.: Analysis and detection of spam accounts in social networks. In: 2016 2nd IEEE International Conference on Computer and Communications (ICCC). IEEE (2016)
- [5] A. Karami and L. Zhou, "Improving static SMS spam detection by using new content-based features," vol. 51, no. 1, pp. 1–9, 2014.
- [6] X. Jin, C.X. Lin, J. Luo, and J. Han. Socialspamguard: A data mining-based spam detection system for social media networks. *PVLDB*, pp 1458–1461, 2011.
- [7] A. Karami and L. Zhou, "Improving static SMS spam detection by using new content-based features," *Improving static SMS spam detection by using new content-based features*, vol. 51, no. 1, pp. 1–9, 2014.
- [8] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk E-mail. In *Learning for Text Categorization – Papers from the AAAI Workshop*, pages 55–62, Madison, Wisconsin, 1998.
- [9] P. Pantel and D. Lin. SpamCop: a spam classification and organization program. In *Learning for Text Categorization – Papers from the AAAI Workshop*, pp 95–98, Madison, Wisconsin, 1998.
- [10] A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classification. In *AAAI'98 Workshop on Learning for Text Categorization*, pages 41–48, Madison, Wisconsin, 1998
- [11] Jin, Z. (2008). Spam message self-adaptive filtering system based on Naive Bayes and support vector Machine. *Journal of Computer Applications*, 28(3), 714-718. doi:10.3724/sp.j.1087.2008.00714
- [12] M. Hossain, I. Jui, and A. Suzana, "Sentiment analysis for Bengali newspaper headlines", *Hdl.handle.net*, 2019.
- [13] T. A. Almeida, T. P. Silva, I. Santos, and J. M. G. Hidalgo, "Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering," *Knowledge-Based Systems*, vol. 108, pp. 25–32, 2016.

- [14] "Naive Bayes text classification", Nlp.stanford.edu, 2019. [Online]. Available: <https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html>. [Accessed: 02- Apr- 2019].
- [15] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," in *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, 1998, pp. 41--48.

Malicios Bangla Text Content Online

ORIGINALITY REPORT

6%

SIMILARITY INDEX

3%

INTERNET SOURCES

0%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	dspace.library.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	scholar.sun.ac.za Internet Source	1%
4	open.library.ubc.ca Internet Source	<1%
5	www.math.unipd.it Internet Source	<1%

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off