BANGLA GURUCHANDALI DOSH SENTENCE DETECTION USING MACHINE LEARNING TECHNIQUES

BY

Rozanee Kanta Das
ID: 181-15-11126

Alaya Refat Tinni
ID: 181-15-11128

Tanjina Zaman Rinvee
ID: 181-15-10964

This Report Presented in Partial Fulfillment of the Requirements for

The Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Sadekur Rahman

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By

Zerin Nasrin Tumpa

Lecturer

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Project/internship titled “**Bangla Guruchandali Dosh Sentence Detection Using Machine Learning Techniques**”, submitted by **Rozanee Kanta Das**, ID No: 181-15-11126, **Alaya Refat Tinni**, ID No: 181-15-11128, **Tanjina Zaman Rinvee**, ID No: 181-15-10964 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 04-01-2022.

BOARD OF EXAMINERS

Chairman

____

Dr. S.M Aminul Haque (SMAH)
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Tania Khatun

Internal Examiner

____

Tania Khatun (TK)
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

____

Md. Sazzadur Ahamed (SZ)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

____

Dr. Shamim H Ripon
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We here by declare that, this thesis has been done by us under the supervision of **Md. Sadekur Rahman, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Sadekur

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Zerin Nasrin Tumpa
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Rozanee Kanta Das

Rozanee Kanta Das
ID: 181-15-11126
Department of CSE
Daffodil International University

Alaya Refat Tinni

Alaya Refat Tinni
ID: 181-15-11128
Department of CSE
Daffodil International University

Tanjina Zaman Rinvee

Tanjina Zaman Rinvee
ID: 181-15-10964
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

At first, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final thesis successfully.

We really very grateful and wish our profound our indebtedness to **Md. Sadekur Rahman**, Assistant Professor, Department of CSE Daffodil International University, Dhaka. Deep knowledge and keen interest of our supervisor in the field of "Natural Language Processing and Machine Learning" to carry out this thesis. His scholarly guidance, patience, constructive criticism, motivation, constant and energetic supervision, continual encouragement, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this thesis.

We would also like to thank our co-supervisor **Zerin Nasrin Tumpa**, Lecturer, Department of CSE Daffodil International University, Dhaka. When we face any problem, she helped us with valuable ideas and suggestions. She motivated us and help us to complete this work.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan, Professor & Head,** Department of CSE, for his motivation and appreciation. We are also very thankful to other faculty members 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 are very thankful to our parents and friends who were always motivate and criticize our work in a manner to improve our work. At least we thank all of them from the core of our heart.

ABSTRACT

Our life is surrounded with technology and we can’t live without this technology. Technology is upgrade day by day. Using Natural Language Processing (NLP) techniques computer can understand human language. Now a days, by the help of NLP researcher are interested to work with text document classification. Bangla text document classification, sentiment analysis etc. are interested topic for researcher. So, in our work we are going classify Guruchandali Dosh of Bangla sentences. In our Bangla language peoples are familiar with Sadhu and Colito form. Colito form is uses in our daily life and Sadhu form is used to written Bangla literature, novel, poems etc. When two forms of Bangla language mixed up in a sentence this is called Guruchandali Dosh. We our work we are going to detect the Guruchandali Dosh sentences using supervised learning techniques. In NLP work text document are easy to preprocess and translate. So, we collect Sadhu and Colito form of data from various Bangla text book, novel, poems and newspaper. Then we make our dataset changing the sentences using some Bangla grammatical rules. Finally, we are able to collects 1712 Bangla text data. We need to preprocess our data before using the machine learning algorithms. We preprocessed our text raw data by removing unwanted data, Stop Words etc. After that we use six classification techniques to classify Guruchandali Dosh sentences. In our work we use Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Extreme Gradient Boosting (XGB), Support Vector Machine (SVM), K-nearest neighbors (KNN) algorithms. All algorithms perform very well on our datasets. Among them Multinomial Naive Bayes (MNB) algorithm came with highest accuracy which is 85%. When we give input Bangla text data in our model, MNB model is able to predict the Guruchandali Dosh perfectly.

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTER

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Using Natural Language Processing (NLP) techniques, it become too easy to communicate between human language and computer. Because it helps to analyze the text automatically and then categorized it into the related contents called text classification. A vast number of research work has been done already on different Language by text classification specially the binary text classification one but there is a small number of works has done on Bangla language where Bangla language takes the 7th position among 100 of others language. [1] So, it shouldn't be contaminated. But it become contaminated when a sentence loss it's compatibility (যোগ্যতা) which means terms of semantic (অর্থসংক্রান্ত) and ideological (ভাবগত) consistency in a sentence. One of the faults of compatibility is Guruchandali Dosh (গুরুচন্ডালী দোষ). It occurs when linguistic error happened. Mainly the defective mixture of Colito Bhasha and Sadhu Bhasha makes Guruchandali Dosh. Sadhu Bhasha is come from tathsomo (তৎসম) shobdho which is directly come from Sanskrito (সংস্কৃত) when the Bangla language was born. It is once used for writing activities such as stories, poems, drama or in official work like application, letter, newspaper and so on. Whereas Colito Bhasha was verbal language mainly but now it also uses in writing. Now people granted it more than Sadhu bhasha because it is easier. But people are mixing up these two forms and make Guruchandali Dosh continuously which makes the language inconsistent and impure. sometimes people even get confuse if a sentence is really Sadhu or Colito. [2]

For Example: আমার যা ইচ্ছা (সাধু) আমি তা (চলিত) করব। (I will do what I want to do)

Most of the people will think that there is no Guruchandali Dosh in the above example but there is. Because "ইচ্ছা" is the Sadhu form and "তা" is the Colito form. That is why we are going to construct a model by machine language that will prevent of making this type of mistake and confusion and keep pure our language.

For our model, text classification is needed which will classify texts from any topic. It will peruse the text automatically by Natural Language Processing (NLP) and categorized the text on the base of dataset. Among the variety of text classification in NLP, binary and multi-level text classification are most common. Comparatively Binary text classification is the easier one. That is why we will work mostly with binary classification in this paper. As its characteristics is to detect a portion of the text sentence so if the selected portion is in categorized data or not, so that we have two classes in our dataset. One is Guruchandali Dosh and another one is Guruchandali Dosh Nei. But it will take less time as we are using Natural Language Processing (NLP) where it will automatically generated text. It also helps to reduce charge and manpower. Moreover, it gives the accuracy most perfectly. But for accurate result we need a huge dataset that can cooperate the model for better procedure, form and requirements. That is why dataset play a significant role so for we need to gather proper data without garbage though for unnecessary data in the model, there has a cleaning process. But we need pure dataset to have accurate outcome because there will be comparison between training and input text. For this process first we need to give an input for text classification and the system will compare with training text that will already save in the model and then categorized it to the nearest one.

In our work, six classifiers are used for an identical output. They are: Decision Tress (DT), Naive Bayes (NB), Random Forests (RF), Extreme Gradient Boosting (XGB), Support Vector Classifier (SVC) and K-nearest neighbors (KNN). We will narrate the whole process in Chapter3 – Research Methodology.

1.2 Motivation

Natural Language Processing is the way to build the communication between human and machine. Now a days using Natural Language Processing techniques there are lots of AI was build which is able to talk human language. Beside this in English language there are many tools able available which is able to correct the grammatical fault of English language. But in Bengali language we cannot find that type of tools. Our native language is Bengali. So, if we are not contributing in Bengali NLP domain, then this machine programmed tools will not be developed in our language. From this motivation we selected Bengali NLP topic for our research work. We select this Guruchandali Dosh sentence detection topic, because sentence fault detector help people to correct their Bengali sentence. When people type Bengali language, they made mistake by mixing up two

Bengali from Sadhu and Colito. So, from this motivation we decided to build a model which can detect the Bengali sentence error perfectly.

1.3 Rationale of the Study

In this age of modern technology, modern tools are update day by day. If we talk about sentence auto correction tools, we have so many tools available in English Language. But we cannot find any sentence auto correction tools for Bengali Language. When we type some official documents in Bengali Language, we made some mistake. Sometime we mixed up Sadhu and Colito form. Grammatically it is an error. But we cannot find this, because we don’t have any auto correction tools in Bengali Language. If we know about the technology this auto correction tools made up using NLP techniques. Bengali is our native language if we can’t work with this language no one can do. That’s why there is no auto correction tools available for Bengali. So, this is a problem in our modern age of technology. That’s why we create a dataset and proposed a model to detect Bengali Guruchandali Dosh sentence.

1.4 Research Questions

During the research work some question occurs about this work. The main questions of our work in given below:

- How to collect and preprocess Bengali text data?
- How to extract features from Bengali dataset?
- Which classifiers performs better to detect Guruchandali Dosh from Bengali sentence?

1.5 Expected Output

This is our experimental project, but our main target is to make a paper about this project. We found so many related works about sentence fault detection. But most of them are in English language. We also found some Bengali sentence fault detection, but there is no work found about Guruchandali Dosh detection. So, we decide to try some experiment about this topic and willing to make a model using machine learning techniques which can detect the Guruchandali Dosh in Bengali Sentence. For this reason, we decided to make a dataset of this project. In the end we have some expected output. Outcome are given below:

- Make a more accurate system
- Published one or more papers on International Conference
- Make a model for sentence autocorrection tools in Bengali Language
- Make a Bengali Guruchandali Dosh dataset
- Our main expected output is our system can detect the Guruchandali Dosh accurately

1.6 Research Layout

In our report we have total 6 chapters

- In Chapter 1 we mention our whole research work’s outline and divided this chapter into multiple subchapters. For example, introduction, motivation, rational of the study, research question and expected output of our project.
- In Chapter 2 we have discussed about the previous work on Bengali text classification, the scope of the problem and challenges in this work.
- In Chapter 3 we will talk about our work procedure, methods and techniques to build a Bengali Guruchandali Dosh detector model.
- In Chapter 4 we will discuss about the Experimental Results and Discussion of our build model.
- In Chapter 5 we will talk about the Impact of Society, Environment, Ethical Aspects and Sustainability plan of our work.
- In Chapter 6 we have discussed about the Summary, Conclusion and Further Study of the work.

CHAPTER 2

BACKGROUND

2.1 Terminologies

Recent years natural language processing has been a hot topic for researchers. Maximum work was done in the English language. But nowadays there are more NLP model builds based on the Bangla text datasets. Such as Sentiment analysis from the Bangla dataset, Bangla Fake news detection, Bangla Text Classification, Bangla Saint (Sadhu) and Common (Colito) form classification and Bangla Text Summarization. But there is no dataset available for Bangla Guruchandali Dosh (A sentence which is mixed up with Saint and Common form). [18] Our work is related to Bangla core linguistics. We found a few papers related to our work. In our work we are going to classify Bangla Guruchandali Dosh sentences. To implement our work, we need to introduce with some new terminology. Working with NLP is not an easy task. In NLP work we only collect raw text data, but to apply algorithms we need to convert this into numeric value. For this we introduce new term Countvectorizer and TF-IDF vectorizer. As well we will mention some term like k-fold cross validation, confusion matrix, classification report, hyperparameter tunning. In the upcoming chapter we will discuss about this term elaborately. To implement our work perfectly and to know about this new term we reviewed some previous work which is related to Bangla Natural Language Processing. Some of them are shortly described below.

2.2 Related Works

A research paper was published in 2020, about Bangla Saint and Common form classification. In their work they proposed a method to classify Bangla Saint (Sadhu) and Common (Colito) form. In their work they collect 1200 data from various Bangla text books, newspapers and online blogs. Then they preprocess their data by removing special characters and stop words. Before using machine learning algorithms, they convert the text document data into vectors using Countvectorizer technique. They use multiple machine learning classification algorithms such as DT, NB, XGB, KNN, RF and SVC. Naive Bayes perform very well in their datasets with accuracy 77%. [3]

In 2020, a research paper was published about sentiment analysis from the Bangla depression dataset. They collect data from various social media sites and Bangla blogs. They process their data using Bangla text processing techniques and they tokenize the data using Countvectorizer. In their work they have two different classes. They use some classification algorithms to predict people's sentiment. They use six different algorithms such as KNN, MNB, SVC, DT, RF and XGB. Among all of the algorithms, Multinomial Naive Bayes gave the highest accuracy for their datasets which is 86.67%. [4]

Hussain, Mahmud et al [5], applying different machine learning algorithms to detect Bangla abusing text. They collect data from social media and preprocess that data using Bangla text processing. They proposed root level algorithms and unigram string features to detect abusing Bangla text. They make 3 sets of comments as training data. In every set they increase the number of comments. When the number of comments increases, the accuracy increases 20% from the previous state.

Haque, Mridha et al [6], proposed a method to identify Bangla grave and extreme guilt faults. They collect Bangla text data from books and newspapers. After collecting the data, they preprocess the data by removing special characters. They tokenize the data using Countvectorizer. They use different machine learning algorithms like LR and SVM. For extreme guilt and grave fault SVM algorithms perform very well and give the highest accuracy which is 87% for extreme guilt fault and 85% for grave fault.

Mamun and Shahin [7], proposed a model to detect social media Bangla bullying text. They collect data from social media sites such as Facebook and Twitter using API. They collect 2400 text data. After collecting the data, clean and preprocess the data using some preprocessing techniques. They proposed a method to classify the Bangla bullying text. For this reason, they use classification algorithms. In their work they used NB, J48, SVM and KNN (1-nearest neighbor) and KNN (3-nearest neighbors). Among all algorithms, the Support Vector Machine algorithm performs very well with performance accuracy 97%.

Hussain, Hasan et al [8], published an experimental analysis to detect Bangla fake news using supervised learning algorithms such as SVM and MNB. In their work they collect fake and real Bangla news data from various Facebook, YouTube and other social media sites. In their dataset the ratio of real and fake news was 60% and 40%. They preprocess their data by removing special characters, punctuation marks, numerical values and emojis. After that they extract the feature from the data using Countvectorizer and TF-IDF vectorizer. They use two supervised machine learning algorithms, SVM and MNB. SVM algorithm scores remarkable accuracy very well with their Bangla fake news dataset which was 96%.

Alberto, Claudio et al [9] suggest a method to automated binary text classification using machine learning. Text classification techniques are recently most popular for data analysis. Digital format of text data is increasing day by day. Enormous increase of text data we need to organize the data. If this task can be done automatically that will reduce time. In their work they proposed a technique to classify binary text automatically. For their work they use Spanish and English datasets. They apply some machine learning algorithms such as DT, SVM, NB for automated binary text classification in two different language datasets. But there is no significant difference for binary classification in two different languages. They showed that SVM and NB perform pretty well for binary text classification. In their work they got 97% accuracy in SVM and 90% accuracy in NB algorithm.

Kim, Han et al [10] suggest an effective method for classification algorithm naive bayes text classification. For their work they use two commonly used corpus. In one corpus there are 21578 news articles data and another corpus 19997 articles. For feature extraction from text, they use feature weights. In their work they proposed poisson naive bayes text classification model with feature weights model. They use pre-document term frequency normalization for converting the text into vectors. For their used dataset poisson naive bayes classifier achieves score 0.90 for first corpus and 0.86 for second corpus.

Ashis and Rikta [11] proposed methods for categorizing Bangla web text documents using supervised machine learning methods. In their work they use four supervised machine learning algorithms such as KNN, DT, SVM and NB. They use Bangla newspaper corpus. They make their own corpus with 1000 text documents. Each document belongs to five categories. They preprocess the data by removing punctuation marks, special characters, numerical digits and stop words. They extract the feature from the text using TF-IDF vectorizer techniques. They split their data into two parts. For training purpose, they kept 80% and for testing purpose they kept 20% data. After that they apply four supervised machine learning algorithms. Among four classification algorithms SVM algorithm shows decent accuracy for their dataset which is 89.14%.

2.3 Comparative Analysis and Summery

For our work we reviewed some previous work related with us. We are working with Bangla NLP, that's why we related some paper which is related with Bangla test data. Basically, we are going to check which machine learning algorithms perform very well with Bangla text data. In this section we are going to compare one previous with other work. In Table 2.1 shows the comparison between previous Bangla NLP work.

Table 2.1: Comparison Between Bangla NLP Previous Work

Work Title	Work Type	Best Algorithms Name	Best Accuracy Score
Bangla saint and common form classification	Binary Classification	Multinomial Naive Bayes	77%
Sentiment analysis from the Bangla depress dataset	Multi-Class Classification	Multinomial Naive Bayes	86.67%
Identify Bangla extreme guilt fault	Binary Classification	Support Vector Machine	87%
Detect social media Bangla bullying text	Binary Classification	Support Vector Machine	97%
Detect Bangla fake news	Binary Classification	Support Vector Machine	96%
Categorizing Bangla web text documents	Multi-Class Classification	Support Vector Machine	89.14%

We reviewed many previous works which is related with text classification problem. But in the above table we only mention the work with is related with Bangla text classification. Because that will be very helpful for us to compare our proposed model. From the above table we saw that most of the work is related with Bangla NLP and Bangla Text classification. Most of the case Support Vector Machine classifier perform very well with text data. So, from the previous work we got an initial and primary idea for our work that which algorithms we use for our work.

2.4 Scope of the Problem

When we read and analysis the previous Bangla NLP work, we found some work like sentiment analysis, Bangla text classification, Fake new detection etc. We found one work which is about Bangla sentence fault identification. They identify extreme guilt fault in Bangla it’s mean is Bahullo Dosh (বাহুল্য দোষ). But we can’t find any work related with detect Guruchandali Dosh sentence. We find a work which only detect Sadhu and Colito Bhasha. So, we decided to work with this topic. Because Guruchandali Dosh is major problem for Bangla Language. Being a Bangladeshi citizen, our mother tongue is Bengali. We all are comfortable and fluent in Bengali Language. So, we need to develop our technology for our Bengali Language. Compare with English Language there are lots off work available related with English sentence correction. And also, there are lots of dataset and tools available for English Language. But in Bengali Language few numbers of dataset and tools are available. So, from Bangladesh perspective we need to work with our Bengali Language. Because we need to write our official and government document in Bengali Language. But all of the document writing tools is not integrated Bengali auto correction tools. For build this tool we need dataset and machine learning model which can detect and correct our sentence fault. In Bengali Language a common mistake Guruchandali Dosh occurs. But we can’t find any model and tools which can detect Guruchandali Dosh in Bengali Sentences. That’s why we decided to make a model which can capable of detect Guruchandali Dosh.

2.5 Challenges

During the whole process of our work, we face some challenges. The main and first problem is dataset. Finding Bengali data is not an easy work for a researcher. If someone found some data luckily, but they are not structure or properly encoded. In our case we faced same problem. We collect some Bangla text book but unfortunately some books are encoded with some techniques we can’t decode that. After trying so hard we managed to collect some data. But we found correct data. There was no dataset available for Guruchandali Dosh. We read some blog and books about Bangla grammar and make a Guruchandali Dosh dataset. After data collection in preprocessed stage, we face some more problem. Because there is no built-in library available for Bengali data preprocessed. We need to use regular expression to remove system and extra unwanted character. We also need to make a stop words dataset for remove stop words. We need to modify some stop words data for our work purpose. So, for this limitation and challenges we can’t collect more Guruchandali Dosh data. If the dataset is large, the outcome will be more precise and accurate. But, in the end we are able to make a model to detect Bengali sentence fault.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

In our work we want make a model which is able to detect Bangla Guruchandali Dosh Sentences. To make this model we want to make a dataset first and we need to understand that which type of work we are doing. In our work we have two classes. So, our work is binary classification. In this we are basically solving classification problem. To build a model there are two ways in machine learning algorithms. Supervised Learning and Unsupervised Learning. In Supervised Learning we are given input and output to the system and system predict the unseen data based on the input output data. On the other hand, in Unsupervised Learning, we are given only input data, machine cluster that data based on the data pattern. In our work we are given input and output data to the system to train the model. So here we use supervised learning. In supervised learning there are some classification algorithms exists which is use to solve classification problem. Our work is related to binary classification problem. So, we will use some classification algorithms which is perform and give high accuracy with text data. In our work we are going to use KNN, SVM, MNB, XGB, DT and RF classification algorithms. In the upcoming proposed methodology section, we will discuss about all of the algorithms, how it works and which algorithms perform very well in our dataset. After that we need to evaluate our model based on some criteria like precision, recall and f1-score. We will briefly discuss all terms in the proposed methodology section.

3.2 Data Collection Procedure

Machine Learning algorithms perform very well when the collecting data is more balanced and reliable to train the machine perfectly. So, data collection is very important for our work. For our research work we need to make our own Bengali Guruchandali Dosh (গুরুচন্ডালী দোষ) dataset first. But there is no data available in Bangla literature where Guruchandali Dosh was found. In Bangla literature the authors wrote their fiction or book in Sadhu accent or Colito accent. For this reason, data collection is quite challenging for us. So, finally we decided to make two classes in our dataset. One is Guruchandali Dosh and another one is Guruchandali Dosh Nei. We collect Guruchandali Dosh Nei type sentences by collecting Sadhu and Colito accent sentences from Bangla fiction, Bangla newspaper and several online sources. Then we read some rules about Guruchandali Dosh from Bangla Grammar Books and Blogs. [12] After that we make our Guruchandali Dosh type sentences from our Guruchandali Dosh Nei type sentences. Finally, we are able to collect 1712 data. Table 3.1 provides the sample data.

Table 3.1: Bangla Dataset of Guruchandali Dosh

Sentence	Sentence Type
আমাদের চারপাশের পৃথিবী যে দ্রুত বদলে যাচ্ছে তা সহজে মেনে নেওয়া আজো অনেকের পক্ষেই বেশ শক্ত	গুরুচণ্ডালী দোষ নেই
তৎকালে আমাদিগের পরিধেয় বস্ত্র বলিতে ছিল একখানা পাঁচ হাতের মলিন ধুতি।	গুরুচণ্ডালী দোষ নেই
সে যে পরের ঘরের দাসদাসী এবং কর্তাগৃহিণীদের অনুগ্রহের ওপর নির্ভর করে বাস করিতেছে	গুরুচণ্ডালী দোষ
সে সময়ে আমাদের পরিধেয় বস্ত্র বলিতে ছিল একখানা পাঁচ হাতের মলিন ধুতি	গুরুচণ্ডালী দোষ
কিন্তু ছেলেদের যে গৃহহীন করিতে বসিয়াছেন সে কথা তাহাদের নিকট হইতে গোপন রাখিলেন	গুরুচণ্ডালী দোষ নেই

3.3 Statistical Analysis

After collecting data from various source, we are able to collect 1712 data. In our dataset we have 1712 sentences. 890 belong to the Guruchandali Dosh nei class and the rest 822 belong to the Guruchandali Dosh class. Table 3.2 provides the percentage of two classes data in our dataset. After divided the two class we calculated the length of the each collected sentence. Table 3.3 shows some data with their length. Length helps us to calculate our vocabulary that means it helps us to set the value of max features in Countvectorizer. In the below section we provide the statistical analysis of our dataset.

Table 3.2: Dataset Details Information

Type of Sentences	Total Data Count	Percentage of Total Count
গুরুচণ্ডালী দোষ নেই	890	51.98%
গুরুচণ্ডালী দোষ	822	48.02%

Table 3.3: Sample Dataset with Length

Sentences	Type	Length
আমাদের চারপাশের পৃথিবী যে দ্রুত বদলে যাচ্ছে তা সহজে মেনে নেওয়া আজো অনেকের পক্ষেই বেশ শক্ত	গুরুচণ্ডালী দোষ নেই	**16**
তৎকালে আমাদিগের পরিধেয় বস্ত্র বলিতে ছিল একখানা পাঁচ হাতের মলিন ধুতি।	গুরুচণ্ডালী দোষ নেই	**11**
সে যে পরের ঘরের দাসদাসী এবং কর্তাগৃহিণীদের অনুগ্রহের ওপর নির্ভর করে বাস করিতেছে	গুরুচণ্ডালী দোষ	**13**
কিন্তু ছেলেদের যে গৃহহীন করিতে বসিয়াছেন সে কথা তাহাদের নিকট হইতে গোপন রাখিলেন	গুরুচণ্ডালী দোষ	**13**
আর এসব শুধুই যে চারপাশের গাছপালা, পশুপাখির ওপর ছাপ ফেলছে তা নয়, বদলে দিচ্ছে পৃথিবীর জলবায়ু, মানুষের জীবনযাত্রা	গুরুচণ্ডালী দোষ নেই	**18**
স্বদেশের মঙ্গলের জন্য সমস্ত অকাতরে সহ্য করে টানিয়া লইয়া চললাম।	গুরুচণ্ডালী দোষ	**10**

- We have 3 columns in our dataset.
- In our dataset we have highest 20 length of sentence.
- Our dataset is available in CSV (Comma Separated Value) format which extension is .csv
- 1500 unique words available in our data.

3.4 Proposed Methodology

We are going to discuss about our research methodology in this following section. In our work, we use six supervised machine learning classifiers MNB, DT, RF, KNN, SVM and XGB to classify Guruchandali Dosh (গুরুচন্ডালী দোষ) from Bengali sentences. To apply this classification algorithm, we make our own dataset. Though it is hard to find the appropriate resource for Bangla Language but try our level best to make our work accurate. For this we divided our work into some steps. Figure 3.1 represents the steps of our methodology. We discussed data collection procedure in 3.2 section. Rest of the methodology steps are described below.

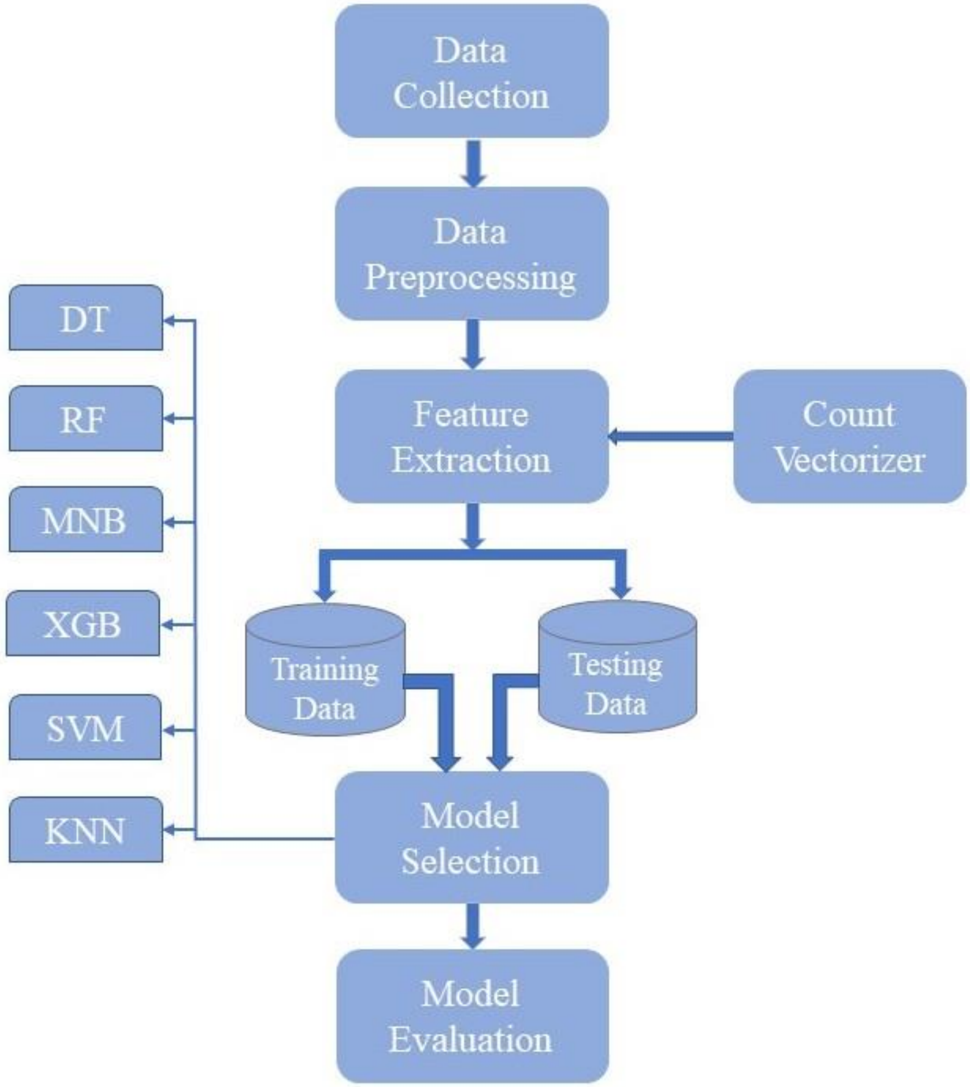


Figure 3.1: Proposed Methodology

3.4.1 Data Preprocessing

We cannot use raw text data to feed our classifier model. Because sometimes raw text data have some characters or symbols which is not essential and suitable for our classifier model. This unwanted characters and symbols sometimes reduce our classifier model accuracy. So, before feeding our model we need to apply some preprocessing techniques on the raw text data. In our raw text data we found some special characters and symbols *, #, !, @ etc. We remove those special characters and symbols from our text. Our text data contains some numerical values such as English digits and Bangla digits. We also found some punctuation marks such as full stop, comma, question marks, quotation marks etc. We use python regular expression library to remove this unwanted data. Table 3.3 shows the characters details which we are remove from our data in preprocessing phase. In our raw text we found some Bangla short form “ইঞ্জিঃ”, “ড.”, “রেজিঃ” etc. We elaborate the short form on Bangla text as “ইঞ্জিঃ” => “ইঞ্জিনিয়ার”, “ড.” => “ডক্টর”, “রেজিঃ” => “রেজিস্ট্রেশন” etc. For this we make a python dictionary with short form and elaborate form, after that we split our text data and compare the data with dictionary. If short form found we replace those data with elaborate form. Table 3.4 provides the details of Bangla probable short form and their elaborate form. In NLP work there are some words in every language which are commonly used and this word are unimportant for machine learning model, this set of data are called stop words. So, in Bengali language we have some stop words such as “ও”, “এবং”, “অতএব”, “অথচ”, “অথবা” etc. [13] In data analysis or when we apply classifier model it creates problems. We need to use Bangla stop words corpus to remove stop words from our dataset. But in our case, before removing the stop word we need to modify the stop words corpus. Because some stop word actually matters for our work, like for example “তাঁর”, it is a stop word but this word can be represented in two Bangla accent (Sadhu and Colito). In Colito accent it is “তাঁর” and in Sadhu accent it is “তাঁহার”. So, we this word in important for our work. Beside this in Bangla stop word corpus we found some words that are important for our work we filter out those word. We remove stop words from our dataset using our own modified Bangla stop words corpus. Table 3.5 shows the raw text data and preprocessed text data and Figure 3.2 shows the text data preprocessing steps.

Table 3.4: Characters Details Considered Removing in Preprocessing

Characters Category	**Characters**
Special Symbols	@, #, $, %, ^, &, *, (,), /, \, {, }, …………
English Digits	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Bangla Digits	০, ১, ২, ৩, ৪, ৫, ৬, ৭, ৮, ৯
Punctuation Marks	?, !, “ ”, ., :, ;, ………..
English Alphabets	A to Z; a to z
Bangla Full Stop (দাঁড়ি)	।

Table 3.5: Bangla Short Form and Elaborate Form

Short Form	Elaborate Form
বি.দ্র.	বিশেষ দ্রষ্টব্য
ড.	ডক্টর
ডা.	ডাক্তার
ইঞ্জিঃ	ইঞ্জিনিয়ার
রেজিঃ	রেজিস্ট্রেশন
মি.	মিস্টার
মু.	মুহাম্মদ
মো.	মোহাম্মদ

Table 3.6: Raw Text Data and Clean Text Data

Raw Text Data	Clean Text Data
শুনিলেন রায়বাহাদুর ঘরে নাই, কিছুক্ষণ অপেক্ষা করিতে হইবে	শুনিলেন রায়বাহাদুর ঘরে কিছুক্ষণ অপেক্ষা করিতে হইবে
বুঝেছিলাম মেয়েটির রূপ বড়ো আশ্চর্য; কিন্তু না দেখিলাম তাহাকে চোখে, না দেখলাম তাহার ছবি, সমস্তই অস্পষ্ট হইয়া রইল।	বুঝেছিলাম মেয়েটির রূপ বড়ো আশ্চর্য দেখিলাম তাহাকে চোখে দেখলাম তাহার ছবি সমস্তই অস্পষ্ট হইয়া রইল
কিন্তু দ্বারিকানাথ তা খেয়াল না করার ভান করে তাড়াতাড়ি পা বাড়ালেন।	দ্বারিকানাথ তা খেয়াল ভান করে তাড়াতাড়ি পা বাড়ালেন
কলে বলিল, “দেখি, মেয়ের বিয়ে দেন কেমন করে।”	সকলে বলিল দেখি মেয়ের বিয়ে কেমন করে
টাকায় ৮ সের সরিষার তৈল আনিলাম	টাকায় সের সরিষার তৈল আনিলাম

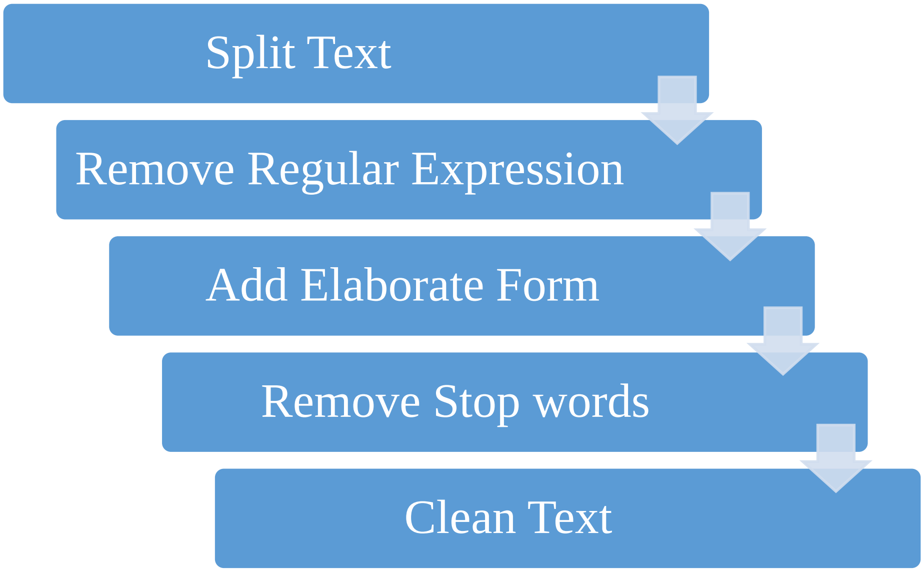


Figure 3.2: Data Preprocessing Steps

3.4.2 Feature extraction

After preprocessed our data we have clean text, but we can’t feed our machine learning model with this text. We need extract the features from clean text. Feature extraction means we need to convert the text data into numerical value. But we need to extract the features in proper way, because extracting proper features from the text have an impact on the machine learning model performance. For this we need to apply some techniques which convert our text data into vector, that means in numerical value. This is called one hot encoding. Here we use count vectorizer and tf-idf (term frequency-inverse document frequency) vectorizer. Count vectorizer is most useful feature extraction method in NLP. Basically, Count vectorizer make a vector from the text data based on the word frequency (count) of each word which is occurs in the sentences. This is very helpful method for sentiment analysis or any NLP related work. Count vectorizer creates a vector or matrix where unique words are represented as matrix columns and each row text data from dataset represented as matrix row. After that it count the word frequency and put that value on the matrix. TF-IDF is advance and common method for features extraction from processed Text data. Sometimes features extraction using TF-IDF vectorizer method increase the proposed model accuracy. Because, it creates a matrix by considering the whole documents of weight of words. The formula of TF-IDF is [14]:

$$tf - idf(t) = tf(t, d) * idf(t) \text{ ----------- } (1)$$

$Where,$

$$tf(t,d) = \sum_{x \in d} fr(x,t)$$

$$idf(t) = log\frac{|D|}{1 + |\{d : t \in d\}|}$$

$Here,$

- ➔ $term\ denoted\ by\ 't' and\ documents\ denoted\ by\ 'd'$
- ➔ $fr(x,t)\ is\ a\ function\ which\ returns\ term\ fequency\ count$
- ➔ $|D|\ is\ the\ number\ of\ whole\ documents$
- ➔ $|\{d : t \in d\}|\ is\ the\ number\ of\ documents\ where\ t\ appers$

In our work we tried two features extraction methods and when we compare which is suitable and increase our model accuracy based on our dataset. We found that Countvectorizer get more accuracy than TF-IDF. So, in our work we use Countvectorizer method to extract feature from our text data. In the upcoming Chapter 4 we will discuss more about this and show the outcome these methods.

3.4.3 Model Selection

In machine learning techniques there are two types of leaning exist. Supervised Learning and Unsupervised Learning. In our work we have input and output data to train a model so we need to use Supervised Learning algorithms. In our work we use some supervised classifier algorithms, we apply six different algorithms DT, RF, XGB, MNB, SVM and KNN on our dataset. After vectorized our data we divided our vector data into two parts training and testing data. We split our data into 80% and 20%. For training purpose, we keep 80% and rest of 20% data for testing purpose. We apply six different classifiers on our training data and evaluate the model based on testing data. Multinomial Naive Bayes classifier perform very well on our dataset. In below, we are shortly discussing about the algorithms and their performance based on our Bangla Guruchandali Dosh dataset.

3.4.3.1 Decision Tree Classifier (DT)

Decision tree is used to solve both regression and classification problems. DT also works with continuous and categorical I/O (input/output) variables. Working approach if this algorithm is to make a tree. DT is a tree where every internal node of the tree represents the attribute values and the leaf node represents the decision. Decision Tree generate high accuracy output for classification problem. In our work it performs very well and score best accuracy which is 78%. The confusion matrix for this classifier is [89 26], [36 104].

3.4.3.2 Random Forest Classifier (RF)

Random forests are a machine learning algorithm which is used to solve regression and classification problems. In these algorithms it split the dataset into many parts and makes many decision trees from datasets. It makes the decision or predict the output based on the decision trees outcome which have the maximum probability of occurrences. For our dataset it predicted the outcome 82% accurate with the confusion matrix [99 16], [24 116].

3.4.3.3 K-Nearest Neighbors (KNN)

K-nearest neighbor is a most used classification algorithm. It is also used in regression problems [15]. It works by calculating the distance between dependent variables (our expected result) and one or more independent variables (our features). For calculating the distance, it uses the Euclidean distance formula. In this algorithm it creates a group by using similar data points that means which data point has closer distance from expected outcome. Based on the value of k (neighbors' numbers) it decides much data it took to create a group. Here we use the value of k as 3. This algorithm cannot predict the outcome, it memorizes the created group and compares the test data with those groups and generates an outcome. For this reason, it takes time to show the expected outcome. That's why this is also known as a non-parametric and lazy algorithm. But for our dataset it performs pretty good with accuracy 70%. The output confusion matrix is [81, 34], [46 94].

3.4.3.4 Support Vector Machine (SVM)

It is a machine learning algorithm which is mostly used to solve regression and classification problems. But it is commonly used in classification problems [14]. In the SVM algorithm the data was plotted in a hyperplane with n-D space (where n is features number). Here we use a two-dimensional surface plane where the line separates the space into two different sections. One class is on one side and another class is on the other side of the space. These algorithms solve our classification problem with medium 79% accuracy. The confusion matrix for this algorithm is [89 26], [27 113].

3.4.3.5 Naive Bayes Classifier (NB)

Naive Bayes is a classification algorithm which is also known as simple probabilistic classifiers. Basically, this classifier is a set of some classification algorithms which works based on Bayes' Theorem. This classifier is not a single algorithm, this is actually a set of familiar classification algorithms where they share a common principle. In our research work, we used Multinomial Naive bayes classifier. Because Multinomial Naive bayes perform very for text document data. Our work is related with text document. This algorithm came with highest accuracy for our dataset. This algorithm perfectly predicted our classes. The accuracy is 83% with confusion matrix [77 134], [38 6].

3.4.3.6 Extreme Gradient Booting Classifier (XGB)

XGB is basically a decision tree-based machine learning algorithm that use the gradient boosting techniques. This algorithm accurate predate the data when the data is unstructured like text data, image data etc. This algorithm widely used for solving regression problem, classification problem etc. This algorithm predicted outcome accurately in our classification problem. This classifier predicted pretty much good accuracy 79% with confusion matrix [100 15], [39 101].

3.4.4 Model Evaluation

Only based on training and testing accuracy we cannot evaluate our model. We need to consider some reports for evaluate our model. First of all, to get accurate result from our model need to apply cross validation. After that we need to make classification report for evaluate our model. The shortly description will discuss in the below subsections.

3.4.4.1 K-Fold Cross Validation

Cross-validation is a validation technique which helps us to evaluate the accurate accuracy of our model. Because when we divide our dataset into train and test data, every time it divides our data randomly. For this reason, sometimes test data consists of data which is not in train data. That's why sometimes we get less accuracy from our model. k-Fold cross-validation helps us to solve this problem. In this technique there is a parameter(k) which is the number of folds that a dataset divided into. Cross-validation randomly divides the dataset into k times and checks how good the model performs when it faces any randomly picked unseen test data. In our research work, we set the value of k as 5. Therefore, we use a 5-Fold cross-validation process in our research work.

3.4.4.2 Classification Report

Only based on the cross validation score we cannot tell that this model is best for this dataset. Besides this we need to evaluate some parameters which are used to make classification reports. These parameters are given below:

3.4.4.2.1 Confusion Matrix

It is a performance measure table which is mostly used to represent the performance of a machine learning model based on a set of test output data [15]. It checks the performance by calculating four terms such as, Ture Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). We will describe briefly about this in the experiment and result segment.

3.4.4.2.2 Precision Score

Precision is a ratio of True Positive result and total Positive predictions. This is also known as PPV or positive predictive value. [16]

$$\text{Precision} = TP / (TP + FP) \text{ ----------- } (2)$$

3.4.4.2.3 Recall Score

It is the quotient of True positive result and the total number of actual predictions. Recall is known as true positive rate or sensitivity. [16]

Recall = TP / (TP + FN) ----------- (3)

3.4.4.2.4 F1 Sore

It is also known as F1-measure. Basically, this is called the Fβ-score. Fβ-score is the combination of harmonic mean of precision score and recall score. When β = 1, this is called F1-score. [17]

$$F\beta - score = (1 + \beta^2).\frac{precision.recall}{\beta^2.precision + recall} \text{ ----------- (4)}$$

$$F1 - score = 2.\frac{precision.recall}{\beta^2.precision + recall}$$

3.5 Implementation Requirements

Our research title is “Bangla Guruchandali Dosh Sentence Using Machine Learning Techniques”. Using the Natural Language Processing techniques, we can extract feature from the text data. So, our work is related with Bengali NLP. We collected Bengali data from various source and we want to make a system which can detect the sentence fault in Bangla. To process and evaluate the entire work we need a high configuration Computer setup with GPU and others necessary instrument. In below we mention the all hardware, software and advance tools which we need to complete this work.

Hardware and Software:

- Intel Core i5 8th gen integrated with 8GB ram
- 1 TB Hard Disk
- Google Colab with 12GB GPU and 350GB ram
- High Speed Internet Connection

Advance Libraries and Tools:

- Windows 10
- Python 3.8
- Pandas

- NumPy
- Regular Expression (RE Library)
- NLTK
- Matplotlib
- Scikit-Learn

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

In this section we are going to describe about our model performance which we apply on our Guruchandali Dosh dataset. In our work we use six classification algorithms which we already discussed in the Methodology section. All classification algorithms perform very well but some of them are classify the text data accurately. When we try to apply Machine Learning algorithms in raw data, we found accuracy below the 10%. After that we preprocessed the raw data. In the Methodology section we mentioned the technique we use to preprocess our data. Then we apply six machine learning algorithms in our clean data. Our DT, RF, MNB, XGB, SVM, KNN models came with the accuracy of 78%, 82%, 83%, 79%, 79%, 70% respectively. These classifiers perform very well with text data. Table 4.1 shows the all-models accuracy with their cross-validation accuracy. Only based on the accuracy score we can't consider our model as a perfect model for our dataset. For the better judgment of our model, we generate confusion matrix, classification report and cross validation score. Then we compare the all values for all models, then we finalize our results.

Table 4.1: Model Accuracy Before Parameters Tuning

Model Name	Accuracy	Cross Validation Accuracy
Decision Tree	78%	76%
Random Forest	82%	81%
Multinomial Naive Bayes	83%	83%
Extreme Gradient Boosting	79%	78%
Support Vector Machine	79%	82%
K-Nearest Neighbors	70%	70%

Among the all-algorithms **Multinomial Naive Bayes** gives us highest accuracy which is **83%**.

4.2 Experimental Results and Analysis

Machine Learning models cannot predict everything with 100% accuracy. We need to try to get an optimal solution from our models. That’s why in our work we experiment our models using some techniques like hyper parameters tunning. Hyper parameters tunning sometimes helps us to find the appropriate models’ parameters for our dataset. So, we decided to tune our models’ parameters based on our dataset. Then use hyper parameters tuning with help of python library. Here we use GridSearchCV to tuning our models. After tunning our models based on our dataset, we see that some algorithms accuracy was increase slightly and some of them are decrease. Our DT, RF, MNB, XGB, SVM, KNN models came with the accuracy score of 73%, 53%, 85%, 79%, 83%, 71% respectively after tuning our models. We see that in XGB classifier there is no change before and after parameters tuning. Also, DT and RF model accuracy decrease after hyper parameters tunning. So, we keep the default parameters for DT, RF and XGB classifier. In the table 4.2 represents the parameters description of all models we used in our work. Hyper parameters tuning experiment came with a good accuracy in our work. But only based on the accuracy score we cannot which model perform very well with our dataset. We need to check our model with unseen data. For this we use k-fold cross validation. In k-fold cross validation dataset divided into k-fold. In the Methodology section we discussed it briefly. In our work we use 5-flod cross validation. After cross validation we got a stable accuracy from models. DT, RF, MNB, XGB, SVM, KNN algorithms came with a good accuracy score of 76%, 81%, 85%, 79%, 83%, 71% respectively.

We also evaluate our model based-on some criteria such as precision score, recall score and f1-score. First, we generate confusion matrix from our dataset using all classification models we mentioned. In the figure represents the visualization of confusion matrix for all models. After that using the confusion matrix, we make classification report for our dataset. We calculate the precision score, recall score and f1-score using confusion matrix. Figure 4.1 shows the confusion matrix for all algorithms we used in our work. Table 4.3 shows the classification report for our all models. In Table 4.4 and Table 4.5 shows the actual result and our used models prediction result for Guruchandali Dosh and Guruchandali Dosh Nei Bangla sentences. All classification algorithms predict expected result for the input bangle text. In the end, after the analysis of our models based on the all-evaluation criteria, we found that Multinomial Naive Bayes (MNB) came out with highest accuracy for our dataset which is 85%.

Table 4.2: Model Parameters Information

Model Name	**DT**	**RF**	**MNB**	**XGB**	**SVM**	**KNN**
Best Parameters	default	default	alpha = 0.9, fit_prior = False	default	C = 25, gamma = 0.01, kernel = ‘rbf’	Metric = ‘manhattan’, N_neighbors = 3, Weights = ‘distance’
Accuracy	0.76	0.81	0.85	0.79	0.83	0.71

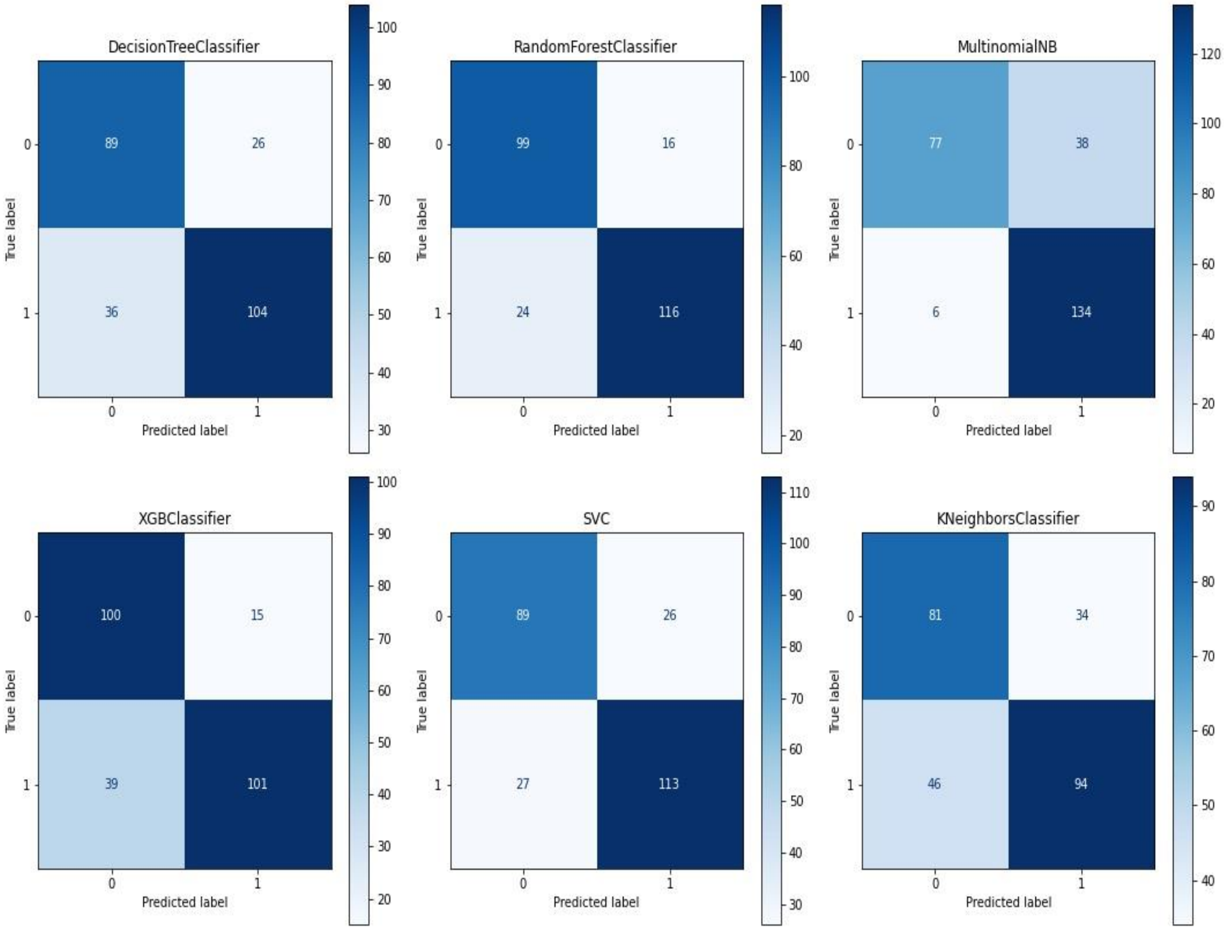


Figure 4.1: Confusion Matrix of All Models

Table 4.3: Classification Report

Algorithms Name	Class	Precision Score	Recall Score	F1 Score	Accuracy Score	Cross Validation Score
Decision tree	Guruchandali Dosh	0.74	0.82	0.78	0.78	0.76
	Guruchandali Dosh Nei	0.84	0.76	0.80		
Random Forest	Guruchandali Dosh	0.79	0.81	0.80	0.82	0.81
	Guruchandali Dosh Nei	0.84	0.83	0.83		
Multinomial Naive Bayes	Guruchandali Dosh	0.91	0.70	0.79	0.83	0.85
	Guruchandali Dosh Nei	0.79	0.94	0.86		
XGB	Guruchandali Dosh	0.72	0.87	0.79	0.79	0.79
	Guruchandali Dosh Nei	0.87	0.72	0.79		
Support vector machine	Guruchandali Dosh	0.77	0.77	0.77	0.79	0.83
	Guruchandali Dosh Nei	0.81	0.81	0.81		
K-nearest neighbors	Guruchandali Dosh	0.68	0.63	0.66	0.70	0.71
	Guruchandali Dosh Nei	0.72	0.76	0.74		

Table 4.4: Original Output and Algorithms Predicted Output

Original Text	**জানি না তাহা সত্যযুগের পল্লিগ্রামে ছিল কি না, কিন্তু একালে তো কোথাও দেখিয়াছি বলিয়া মনে করিতে পারি না**
Original Output	গুরুচণ্ডালী দোষ
Input Text	**জানি না তাহা সত্যযুগের পল্লিগ্রামে ছিল কি না, কিন্তু একালে তো কোথাও দেখিয়াছি বলিয়া মনে করিতে পারি না**
	Algorithms Predicted Output
DT	গুরুচণ্ডালী দোষ
RF	গুরুচণ্ডালী দোষ
MNB	গুরুচণ্ডালী দোষ
XGB	গুরুচণ্ডালী দোষ
SVM	গুরুচণ্ডালী দোষ
KNN	গুরুচণ্ডালী দোষ

Table 4.5: Original Output and Algorithms Predicted Output

Original Text	**কৃত্রিম রাসায়নিক বস্তু ব্যবহারের ফলে উঁচু বায়ুমন্ডলে ওজোন গ্যাসের পরিমাণ কমে যাচ্ছে**
Original Output	গুরুচণ্ডালী দোষ নেই
Input Text	**কৃত্রিম রাসায়নিক বস্তু ব্যবহারের ফলে উঁচু বায়ুমন্ডলে ওজোন গ্যাসের পরিমাণ কমে যাচ্ছে**
	Algorithms Predicted Output
DT	গুরুচণ্ডালী দোষ নেই
RF	গুরুচণ্ডালী দোষ নেই
MNB	গুরুচণ্ডালী দোষ নেই
XGB	গুরুচণ্ডালী দোষ নেই
SVM	গুরুচণ্ডালী দোষ নেই
KNN	গুরুচণ্ডালী দোষ নেই

4.3 Discussion

In the end we are try to contribute in the Bangla research domain. Worldwide there are lots of work in the different language. We are trying to added our Bangla language in that work list. We dreamt to detect the Bangla Guruchandali Dosh sentence. So, we are very happy that we make our dream comes true. In this work we are able to make a machine learning model which is able to detect the Guruchandali Dosh sentences accurately. We use six different algorithms in our model. All model's accuracy is shown in the figure 4.2. Among all of the algorithms Multinomial Naive Bayes detect the sentence's fault accurately with highest accuracy of 85%.

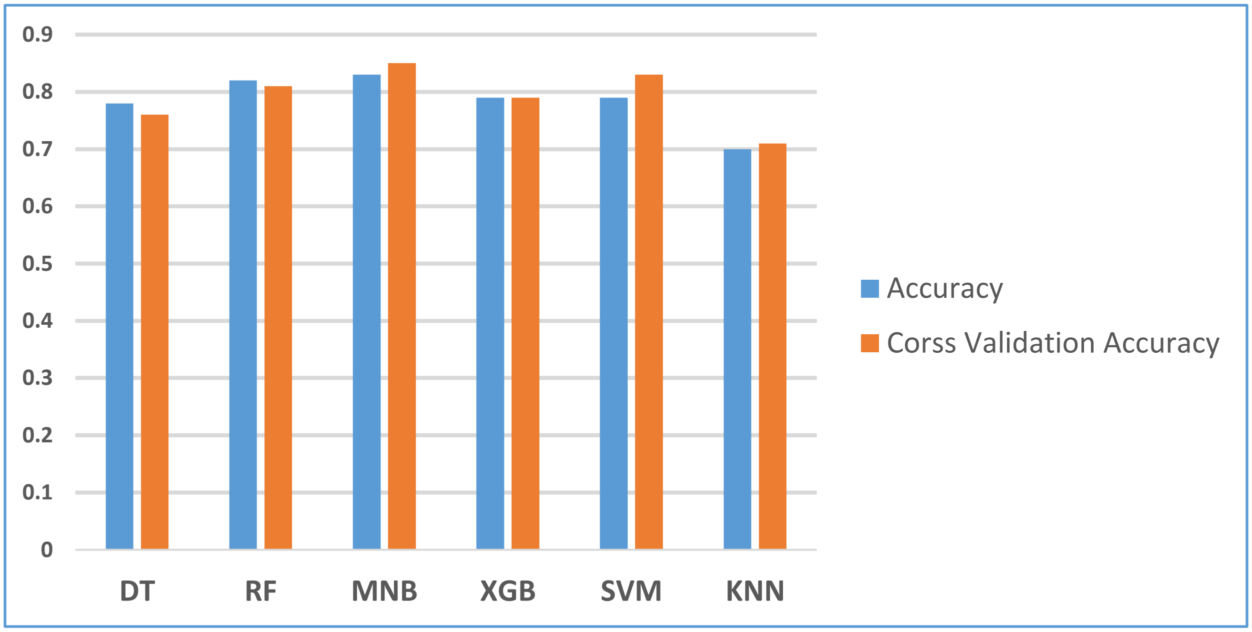


Figure 4.2: Accuracy and Cross Validation Accuracy Bar Chart

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Guruchandali Dosh contaminated our mother language rapidly so it creates a massive impact on Bangla language that is the reason we have created this model so by this anyone can easily have the idea if a sentence has Guruchandali dosh or not. The consequence will spread a great effect in our society. If a society will be constructed of authentic language, then other language's people will have an idea about our dedication for our own language still after achieving it. Others will respect our language as much as their language if we build this kind of society.

5.2 Impact on Environment

As we are expecting that it will be a convenience project so that it will create a great impact on society then it must be having a great impact on environment too. Because an environment is made with a society & society made with people. If people are having adequate knowledge about their own language, then it also dominance in their environment. So, for, our project will help to grow an environment where everyone can say Bangla language or write it in appropriate way without any mistake. Thus, it will build an environment of pure Bangla language.

5.3 Ethical Aspects

It is our moral responsibility to safeguard our own language even after 50 years of acquiring it. That is why we take a petty step to protect our own language. We also have some ethical aspects about our model:

1. To prevent our language from being distort.
2. To protect our language.
3. To help everyone to check the correct form of Bangla language.
4. For enhancing more value of our language.

5.4 Sustainability Plan

Our plan is to help upcoming generation so they can create a society of pure Bangla language & it will be like a chained process that is how it will invent a huge ascendency in our environment as well as society. Guruchandali Dosh is a sentence fault in Bangla Language and it created a massive impact on Bangla language. This is the toughest part in the Bangla grammar. This is the reason why we created the model and why we need to fix this properly so that people can easily find the correct solution. We should have known about this concept and we should learn and clear the concept. That's why we need to sustain this model properly. We need to create a proper model which will provide the correct information about this grammar. This is our mother tongue and we definitely should know these properly. We also need to process how to sustain this model permanently.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary

Our work is related with Bengali NLP. Working with NLP is a challenging work for researcher. In research work dataset is main important for execute the work. In this work we are making a machine learning model for detect Bangla Guruchandali Dosh sentence. This model is very useful for making a programmed Guruchandali Dosh detector. So, for building this model we face some problem and so more things we face during this work. All the steps and work summery is given below step by step.

Step 1: Planning about this work

Step 2: Problem formulation

Step 3: Data collection from various books and newspapers

Step 4: Data Labeling

Step 5: Data cleaning

Step 6: Data Vectorization

Step 7: Train and Test Data Separation

Step 8: Model Selection

Step 10: Model evaluation and performance testing

After execute all of the steps we are finally able to make our model for Detect Bengali sentence fault. Our work contributes in the domain of Bengali NLP research. Working with our native language is pride for us and in the world-wide domain we recognized our language easily.

6.2 Conclusion

There are so many works about Bangla test classification, fake news detection. But Bangla text Guruchandali Dosh detection is a new proposal in Bengali NLP domain. Some classification algorithms in supervised learning perform very well for text data. In our work we proposed to build a model on six different algorithms such as DT, RF, MNB, XGB, KNN and SVM using our Bangla Guruchandali Dosh dataset. In our task making dataset is very hard and tough work for us. Because there is no previous work available related with our data. We need to make our data manually. Finally, in our work we use 1712 data to make a model. All classification algorithms perform very well, but Multinomial Naive Bayes came out with highest performance accuracy with our dataset. Using this algorithms model predict the fault sentences 85% accurately. Therefore, we have some limitation in our work. For the limitation of time and less source of our dataset we are able to collect less numbers of data. If the we are able to collect more data, then our model will be coming out with more accuracy then now. Because ML classifier's accuracy is high when the dataset is large in natural language processing.

6.3 Recommendation

We have some recommendations for our work. In this section we will increase our dataset for improve our model accuracy. In our work we use some supervised machine learning classification algorithms. And in the text data transformation section we use only one vectorization technique. There are so many techniques and algorithms for large number of datasets. So, that model and techniques will predict more accurately Bangla Guruchandali Dosh sentence. Some recommendations of our work are given below.

- Big dataset for Bangla Guruchandali Dosh
- Understand the core Bengali Linguistic to recognize the patterns of Guruchandali dosh
- Understand the transformation techniques and improvised the techniques for Bengali dataset.
- Try to make a better classification model
- Try to get more performance accuracy

6.4 Implication for Further Research

We have some limitations and drawback in our work. For example, we only use machine learning algorithms in our model. Beside this we only use Count Vectorized text transformation techniques. Also, our data is not sufficient. So, we extend our data count. Because, we have plan to apply deep learning algorithms like RNN, LSTM, BiLSTM etc. in our dataset. Without increasing the data volume, we cannot get the better accuracy from deep learning model. We also use Word Embedding techniques to vectorized the text data in to numeric value. In this work we get 85% accuracy but we can’t stop here, we will try our level best in future to get more accuracy from this dataset.

We open our dataset for the NLP researcher. Using this they can also try to get more accuracy. We have a plan to deploy this model on website. Using this website anyone can input the Bengali sentence and get the output that this sentence has Guruchandali Dosh or not.

REFERENCES

[1] Dhaka Tribune, available at <<https://www.dhakatribune.com/world/2020/02/17/bengali-ranked-at-7th-among-100-most-spoken-languages-worldwide>>

[2] Guruchandali Dosh – Gaan Bangla <<https://gaannbangla.blogspot.com/2020/05/what-is-Guruchandali-dos.html>>

[3] Ria, N.J., Khushbu, S.A., Yousuf, M.A., Masum, A.K.M., Abujar, S. and Hossain, S.A., 2020, July. Toward an enhanced Bengali text classification using saint and common form. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.

[4] Khan, M.R.H., Afroz, U.S., Masum, A.K.M., Abujar, S. and Hossain, S.A., 2020, July. Sentiment Analysis from Bengali Depression Dataset using Machine Learning. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE

[5] Hussain, M.G., Al Mahmud, T. and Akthar, W., 2018, December. An approach to detect abusive bangla text. In *2018 International Conference on Innovation in Engineering and Technology (ICIET)* (pp. 1-5). IEEE.

[6] Ul Haque, R., Mridha, M.F., Saha, A.K., Hamid, M.A. and Nur, K., 2020, January. Identification of Extreme Guilt and Grave Fault in Bengali Language. In *Proceedings of the International Conference on Computing Advancements* (pp. 1-5).

[7] Akhter, S., 2018, December. Social media bullying detection using machine learning on Bangla text. In *2018 10th International Conference on Electrical and Computer Engineering (ICECE)* (pp. 385-388). IEEE.Detection of Bangla Fake News using MNB and SVM Classifier

[8] Hussain, M.G., Hasan, M.R., Rahman, M., Protim, J. and Al Hasan, S., 2020, August. Detection of bangla fake news using mnb and svm classifier. In *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE)* (pp. 81-85). IEEE.

[9] Holts, A., Riquelme, C. and Alfaro, R., 2010, November. Automated text binary classification using machine learning approach. In *2010 XXIX International Conference of the Chilean Computer Science Society* (pp. 212-217). IEEE.Some Effective Techniques for Naive Bayes Text Classification

[10] Kim, S.B., Han, K.S., Rim, H.C. and Myaeng, S.H., 2006. Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering*, *18*(11), pp.1457-1466.

[11] Mandal, A.K. and Sen, R., 2014. Supervised learning methods for Bangla web document categorization. *arXiv preprint arXiv:1410.2045*.

[12] Sadhu O Colito Bhasha – Ananya Bangla, available at <<https://ananyabangla.blogspot.com/2020/09/blog-post_1.html >>

[13] Bangla Stop Words List, available at <<https://www.ranks.nl/stopwords/bengali>>

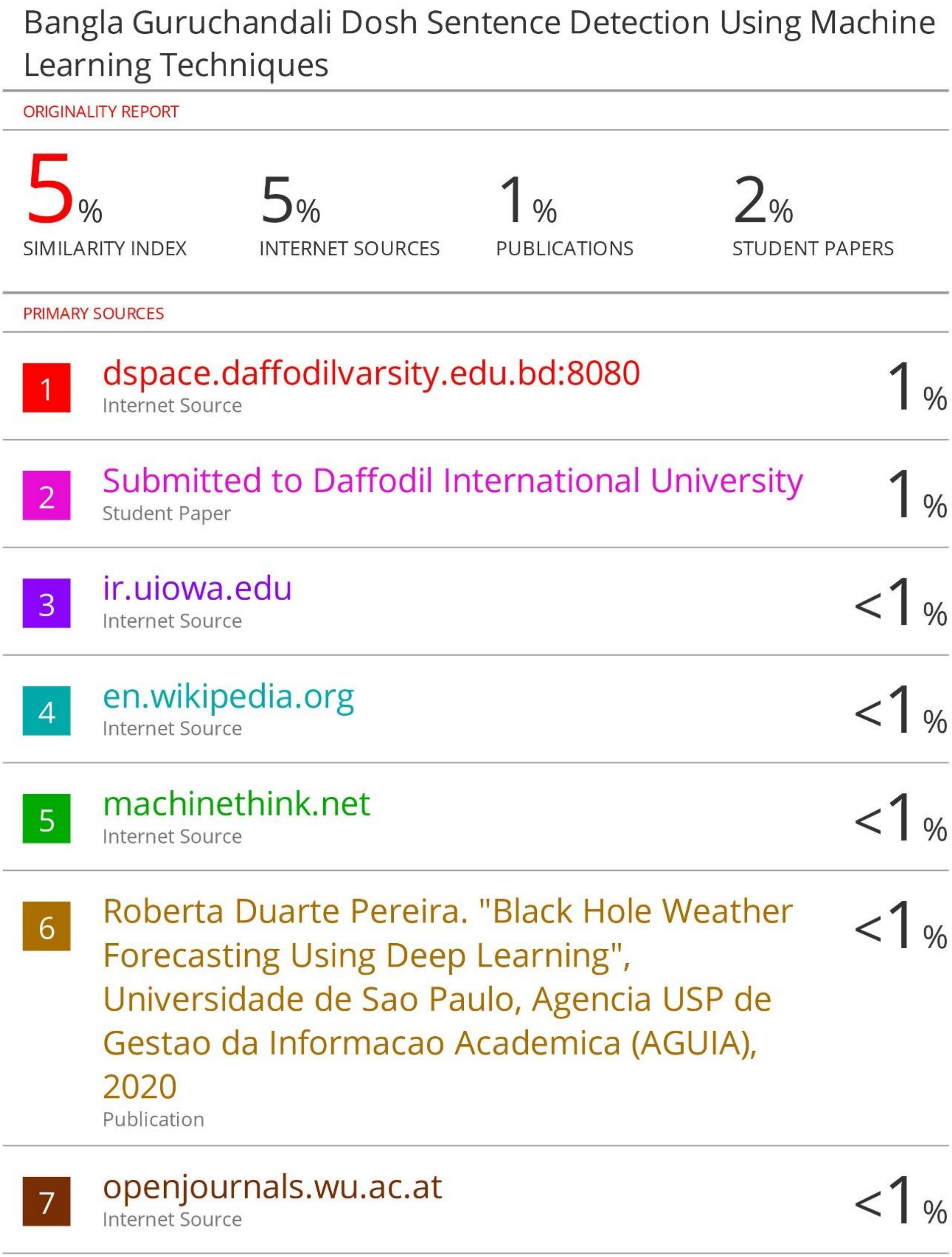
[14] TF-IDF Vectorizer by Mukesh Chaudhary, available at <<https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a>>

[15] 100 Days of ML Coding by Avik-Jain, available at <<https://github.com/Avik-Jain/100-Days-Of-ML-Code?fbclid=IwAR0C1AL320WDL800p3cshofTSfEbZ1tSOW7cmkAUx5X_3QYkPJPqMsxM6mw>>

[16] We found precision and recall details in Wikipedia, available at <<https://en.wikipedia.org/wiki/Precision_and_recall#:~:text=In%20a%20classification%20task%2C%20the,false%20positives%2C%20which%20are%20items >>

[17] Data-School, available at <<https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/#:~:text=A%20confusion%20matrix%20is%20a,the%20true%20values%20are%20known.&text=The%20classifier%20made%20a%20total,the%20presence%20of%20that%20disease)>>

PLAGIARISM REPORT



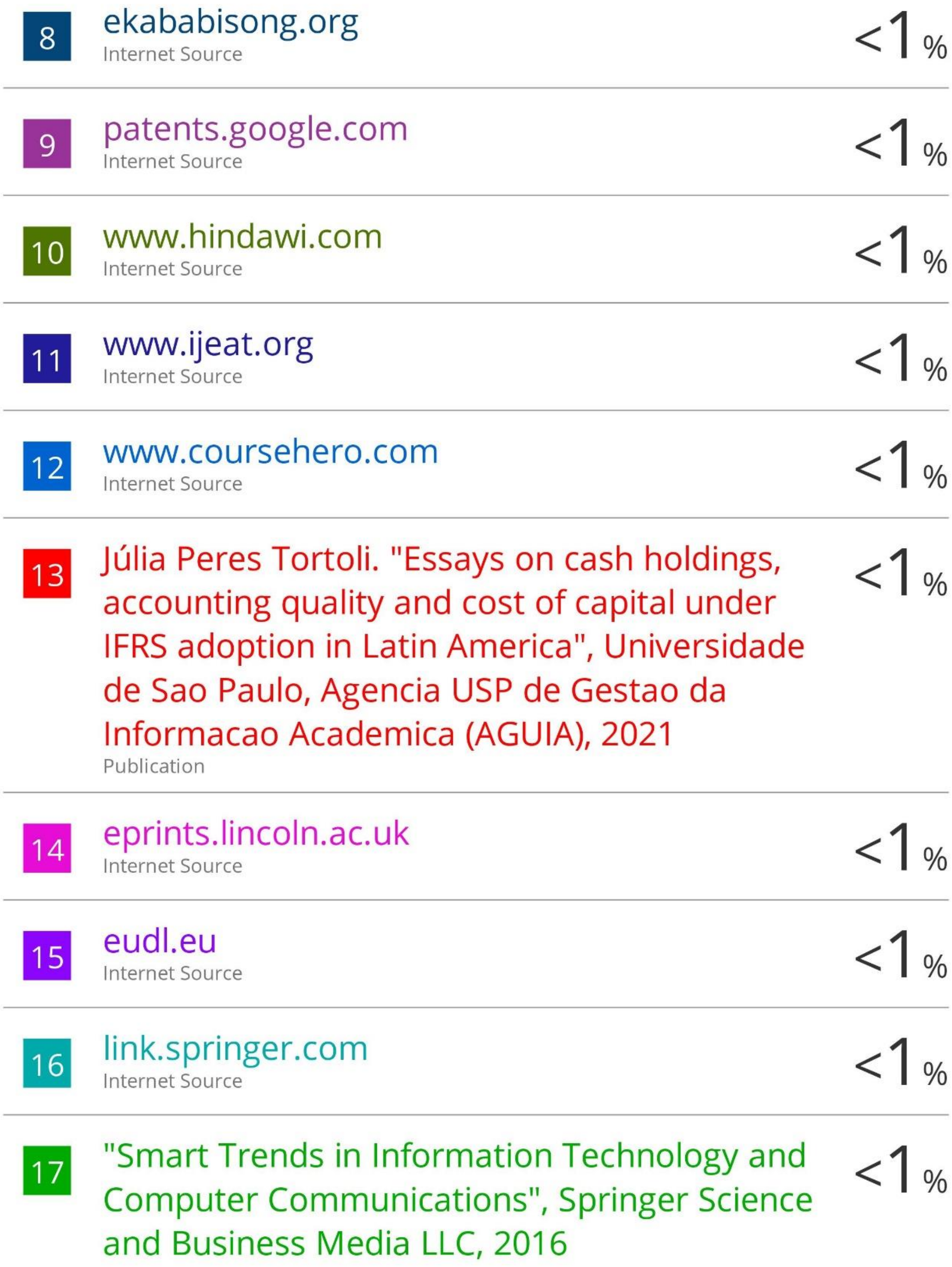
Bangla Guruchandali Dosh Sentence Detection Using Machine Learning Techniques

ORIGINALITY REPORT

5%	5%	1%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
2	Submitted to Daffodil International University Student Paper	1%
3	ir.uiowa.edu Internet Source	<1%
4	en.wikipedia.org Internet Source	<1%
5	machinethink.net Internet Source	<1%
6	Roberta Duarte Pereira. "Black Hole Weather Forecasting Using Deep Learning", Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), 2020 Publication	<1%
7	openjournals.wu.ac.at Internet Source	<1%



8	ekababisong.org Internet Source	<1 %
9	patents.google.com Internet Source	<1 %
10	www.hindawi.com Internet Source	<1 %
11	www.ijeat.org Internet Source	<1 %
12	www.coursehero.com Internet Source	<1 %
13	Júlia Peres Tortoli. "Essays on cash holdings, accounting quality and cost of capital under IFRS adoption in Latin America", Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), 2021 Publication	<1 %
14	eprints.lincoln.ac.uk Internet Source	<1 %
15	eudl.eu Internet Source	<1 %
16	link.springer.com Internet Source	<1 %
17	"Smart Trends in Information Technology and Computer Communications", Springer Science and Business Media LLC, 2016	<1 %

Publication

18	es.scribd.com Internet Source	<1 %
19	tel.archives-ouvertes.fr Internet Source	<1 %
20	www.cs.ust.hk Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		