

BANGLA LANGUAGE MODE (SADHU/CHOLITO)

CLASSIFICATION

BY

Abdul Bari Parves

ID: 151-15-4879

AND

Emranul Haque Rakib

ID: 151-15-5049

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

Supervised By

Md. Riazur Rahman

Senior Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

Zerin Nasrin Tumpa

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2019

APPROVAL

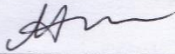
This Project titled “**Bangla Language Mode (Sadhu/Cholito) Classification**”, submitted by Abdul Bari Parves, ID No: 151-15-4879 and Emranul Haque Rakib, ID No: 151-15-5049 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for 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 04/05/2019.

BOARD OF EXAMINERS



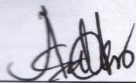
Dr. Syed Akhter Hossain
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



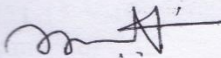
Nazmun Nessa Moon
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Abdus Sattar
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

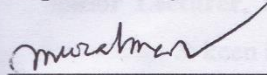
External Examiner

ACKNOWLEDGEMENT

DECLARATION

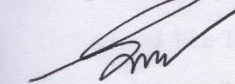
We hereby declare that; this project has been done by us under the supervision of **Md. Riazur Rahman, Senior Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:



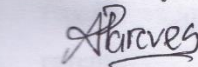
Md. Riazur Rahman
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

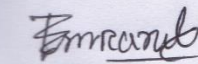


Zerine Nasrin Tumpa
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Abdul Bari Parves
ID: -151-15-4879
Department of CSE
Daffodil International University



Emranul Haque Rakib
ID: -151-15-5049
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

We really grateful and wish our profound our indebtedness to **Md. Riazur Rahman, Senior Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep knowledge & keen interest of our supervisor in the field of “*natural language processing*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

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

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

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

ABSTRACT

This project addresses the problem of distinguishing between two forms of Bangla language, namely Sadhubhasha and Cholitobhasha. The classifier would be beneficial for finding the right word choice for Bangla literature. The main vision of this project is to differentiate the modern era's early Bangla form of Sadhubhasha to the current form of Cholitobhasha. As far as we know there has been no single work done addressing this particular issue. From another perspective, only a few works have been done on "Bangla Language". So, it has been difficult to conduct advanced linguistic works on Bangla language like extracting information or summarizing. We had to face difficulties when collecting Bangla data due to the limited availability, but finally we have collected around total 100000 words dataset for this project. Among which 80% of the data is used for training and rest 20% is test data. Machine learning algorithms Random forest, Naïve Bayes, Support Vector Machine, K-nearest neighbor and Decision tree are applied to classify the language and the Term Frequency-Inverse Document Frequency and Bag of Words are used for the numerical representation. With these classifiers 91% to 99.5% accuracy is observed. The promising outcome of this project is, "sadhu and cholito Language classifier" can be used as the first step on that ladder from where others will be influenced to do further research on Bangla language.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of Examiners	i
Declaration	ii
Acknowledgment	iii
Abstract	iv
CHAPTER	PAGE
CHAPTER 1: INTRODUCTION	1-2
1.1 Introduction	1
1.2 Motivation	1-2
1.3 Research Questions	2
1.4 Expected Outcome.	2
1.5 Layout of the Report	2
CHAPTER 2: BACKGROUND STUDY	3-5
2.1 Introduction	3
2.2 Related Works	3-4
2.3 Research Summary	4-5
2.4 Challenges	5
CHAPTER 3: RESEARCH METHODOLOGY	6-19
3.1 Introduction	6
3.2 Data Collection Procedure	6
3.3 Data Processing	6-10
3.4 Proposed Methodology	10-18
3.5 Statistical Analysis	18

3.6 Implementation Requirements	19
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	20-26
4.1 Introduction	21
4.2 Experimental Result	21-24
4.3 Descriptive Analysis	24-26
4.4 Summary	27
CHAPTER 5: SUMMARY AND CONCLUSION	27-28
5.1 Summary of the Study	27
5.2 Conclusion	27
5.3 Recommendations	27
5.4 Implications for Further Study	27-28
REFERENCES	29

LIST OF FIGURES

FIGURES	PAGE NO
Fig 3.1: stop words for Bangla	9
Fig 3.2: stop words for Bangla	9
Fig 3.3: Random Forest algorithm	12
Fig 3.4: Support vector machine creating hyperplane to classify	15
Fig 3.5: Details of confusion matrix	16
Fig 3.6: Flow Chart of Proposed Model	18
Fig 4.1: List of most used words in Sadhubhasha	24
Fig 4.2: List of most used words in Cholitobhasha	24
Fig 4.3: image plot of most used words in Sadhubhasha	25
Fig 4.4: image plot of most used words in Cholitobhasha	25

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Summary of the related works	4-5
Table 3.1: Format of dataset	7
Table 3.2: Data statistics	8
Table 3.3: Vocabulary Scoring	11
Table 3.5: Sentence binary vector representation	11
Table 4.1: result achieved from random forest algorithm	20
Table 4.2: result achieved from Multinomial Naive Bayes algorithm	21
Table 4.3: result achieved from Gaussian Naive Bayes algorithm	21
Table 4.4: result achieved from Support vector machine algorithm	22
Table 4.5: result achieved from K-nearest neighbor algorithm	22
Table 4.6: result achieved from decision tree algorithm	23

CHAPTER 1

INTRODUCTION

1.1 Introduction

Bengali is sixth most commonly spoken language in the world which was originated and evolved from the Sanskrit in 1000-1200CE. The modern literary form of Bangla language was developed during the 1800 and early 1900 based on the different dialect spoken in the Nadia region, a west-central Bengali dialect. In modern era, Bengali language has two form: Sadhubhasha and Cholitobhasha. Sadhubhasha was considered as the proper form of Bangla language which later took as a form of writing novel and Cholitobhasha is used for the normal conversation. Over the year, rapid changes in tradition and culture might gave us many benefits but also come with some problems like language malformation. Now-a-days, an erroneous desire to become modernized is affecting our Bangla language deleteriously through misusing it. So, people are neither using Sadhubhasha nor Cholitobhasha. Instead, we are becoming habituated to practice malformed Bangla language.

Text classification is known as an important method to handle and process a large number of documents in digital forms which is increasing continuously. Text classification is mainly used for extracting information, text retrieval, and summarization. This project will demonstrate the text classification process through machine learning techniques. In our project, for classifying Sadhubhasha and Cholitobhasha, we have first used Term Frequency-Inverse Document Frequency and Bag of Words model to convert the text document into corresponding numerical features. In addition, Random forest, Naive Bayes, Support Vector Machine, K-nearest neighbor, and Decision tree classifier to classify Sadhubhasha and Cholitobhasha. Our proposed method is expected to perform better than other methods used to classify Bangla text.

1.2 Motivation

Everything evolves with the time which ranges from lifestyle, culture and even language. In modern days, everything is stored in a digital form and its popularity is increasing day by day. As Bangla is our first language, most of the data of our country

is in Bangla. It is inevitable that data is the most powerful source of information. To avail it we need to extract the data, however, most of the tools or methods are created for English and few other popular languages. For this deficiency, most of the data in Bangla could not be extracted and therefore these vast amount of data gets wasted. This current scenario has motivated us to work with our mother tongue Bangla. We have decided to start our work with the earliest two form of Bangla language, Sadhubhasha and Cholitobhasha.

1.3 Research Questions

1. Is it possible to accurately classify the forms of bangla language namely Sadhubhasha and Cholitobhasha?

1.4 Expected Outcome

- To classify Sadhubhasha and Cholitobhasha.
- To find out the most frequently used words in both form of Bengali language.

1.5 Layout of the Report

This report is organized as follows:

- Chapter One includes introduction to the project, motivation, research questions, and expected outcome.
- Chapter Two includes “Background”, related works, research summary, and challenges.
- Chapter Three include Research Methodology.
- Chapter Four includes Experimental Results and Discussion.
- Chapter five includes Summary and Conclusion.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

In this section, we have reviewed some related works, research summary and challenges about our research. In related works section, we will try to explain other research paper and their works, their methods, and accuracy which are related to our work. In research summary section we will give the summary of our related works. In challenges section, we will discuss how we increased our accuracy.

2.2 Related Works

Abu Nowshed Chy, Md. Hanif Seddiqui, Sowmitra Das have proposed a method to classify Bangla news. They have used the Naive Bayes classifier to classify news from news article. They also used RSS crawler for data collection and then build a bangli lexicon and a Bengali stemmer and finally run Naive Bayes classifier [1]. In another project, Andrew McCallum and Kamal Nigam have discussed and compared different model of Naive Bayes classifier. They have compared Multi-variate Bernoulli Model and Multinomial Model. Each of the model perform differently with the variation of data and size of data. In few data set, Bernoulli Model showed good performance, especially in small size data set. In contrast, Multinomial Model performed well with large scale datasets [2]. To classify text, Andronicus A. Akinyelu and Aderemi O. Adewumi have proposed a new method in 2014. As people gets hack every day using phishing email, they have used a machine learning algorithm random forest in their study. The result was impressive and accuracy rate was 99.7 % [3]. Baoxun Xu, Xiufeng Guo, Yunming Ye and Jiefeng Cheng also have proposed an improved method of random forest algorithm for text categorizing earlier. Their proposed feature weighting method and tree selection method an improvement or random forest algorithm. With the new feature weighting method for subspace sampling and tree selection method, they effectively reduce subspace size and improve classification performance without increasing error bound. They have conducted six datasets and all of their proposed Around 70-90% accuracy was achieved by their improved random forest algorithm [4]. Recently, in 2018, Suresh Merugu, M. Chandra Shekhar Reddy, Ekansh Goyal and Lakshay Piplani proposed a supervised machine learning approach

for classifying text messages. They used many supervised algorithms such as SVM, Random Forest, K Nearest Neighbor and BernoulliNB. K Nearest Neighbor performed worst between all of them while Random Forest and BernoulliNB had the best accuracy almost 98% [5]. In another previous study, M. Ikonomakis, S. Kotsiantis and V. Tampakas have discussed about few machine learning techniques for text classification. They stated the detailed information about how a algorithm works, how we should prepare our dataset and how we should preprocess. They also outlined the result evaluation [6]. Timothy P. Jurka, Loren Collingwood, Amber E. Boydston, Emiliano Grossman, and Wouter van Atteveldt discussed about a new tool called RTextTools which are used in text classification for beginners. By using RTextTools one could classify any text only through 10 easy steps. From training to result evaluation by RTextTools are discussed in this paper [7].

2.3 Research Summary

Table 2.1: Summary of the related works

SL	Author	Methodology	Description	Outcome
1.	Abu Nowshed Chy, Md. Hanif Seddiqui, Sowmitra Das	naive Bayes classifier	Classifying Bangla news	78%
2.	Andrew McCallum and Kamal Nigam	Multi-variate Bernoulli Model and Multinomial Model	Comparison between Multi-variate Bernoulli Model and Multinomial Model.	Multinomial Model performed 4.8% better Multi- variate Bernoulli Model
3.	Andronicus A. Akinyelu and Aderemi O. Adewumi	Random Forest	Classifying phishing email from emails.	99.7%
4.	Baoxun Xu, Xiufeng Guo, Yunming Ye and Jiefeng Cheng	Weighting method and tree selection method an improvement	New feature weighting method for subspace sampling and tree selection method, they	70-90% for six different datasets.

		for random forest algorithm.	effectively reduce subspace size and improve classification performance without increasing error bound	
5.	Suresh Merugu, M. Chandra Shekhar Reddy, Ekansh Goyal and Lakshay Piplani	SVM, Random Forest, K Nearest Neighbor and BernoulliNB	Took a dataset of 5000 messages used 90% of them a training and rest for testing. Used different supervised Machine Learning algorithm for classifying.	SVM and Random Forest got accuracy 98% and BernoulliNB 97.6%
6.	M. Ikonomakis, S. Kotsiantis and V. Tampaka	Different Machine learning algorithms	Discussed about different machine learning algorithms and techniques.	Discussed about algorithms.
7.	Timothy P. Jurka, Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, and Wouter van	RTextTools	Discussed about RTextTools by which in ten step one can easily classify text.	Discussed about RTextTools.

2.4 Challenges

The main challenge for our project was not only a huge number of data collection but also make sure that the data is in its purest form. Because we are working on a language's two different form. So, the data we took has to be sadhu and cholito separately.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this section we will discuss about our data collection procedure, data processing, proposed methodology, statistical analysis and implementation requirements. Firstly, in data collection procedure we have discussed how we have collected our data. Next, in data processing part, we have discussed how we pre-processed it for our model. Then in proposed methodology we briefly addressed about the algorithms and methodology that were used for this classification. Consequently, in statistical analysis we highlighted few statistical method and flow charts of the project. Finally, the chapter is closed by a clear concept about what we used for the project.

3.2 Data Collection Procedure

Constant evaluation of culture has a great impact on our Bangla language. Over the year we included many words from many other languages such as English, Hindi, Urdu, Persian, Dutch, Portuguese and many more. At present, we are just not involving new words but also have malformed our language dangerously. Therefore, finding or creating a pure dataset which only has Cholitobhasha or Sadhubhasha are challenging. We were aware of the need for natural raw data in two Bangla forms for the significance of the proposed study. So, we decided to collect data from Bangla novels which were apparently written in Sadhubhasha and Cholitobhasha. Famous bangla literatures of Sarat Chandra Chattopadhyay, Bankim Chandra Chattopadhyay, Syed Mujtaba Ali and Humayun Ahmed were evaluated to form our dataset. We have collected these data from books and different website dedicated for Bangla literature where most of the Bangla writers' book can be found.

3.3 Data Processing

First, we have collected the data in a .docx format. Then, we processed a raw dataset of more than 100000 words. These raw data consisted of Sadhubhasha and Cholitobhasha separately. Consequently, for training we transferred the data to a .xlsx format where we categorized the data into two different classes Sadhubhasha and Cholitobhasha. Every line of the dataset consisted two column text and class.

Data Format and Statistics

Data Format:

Table 3.1: Format of data set

Text	Class
শক্তিশেল বৃকে পড়িবার সময় লক্ষণের মুখের ভাব নিশ্চয় খুব খারাপ হইয়া গিয়াছিল কিন্তু গুরুচরণের চেহারাটা বোধ করি তার চেয়েও মন্দ দেখাইল যখন প্রতুষেই অন্তঃপুর হইতে সংবাদ পৌঁছিল গৃহিণী এইমাত্র নির্বিঘ্নে পঞ্চম কন্যার জন্মদান করিয়াছেন	0
ছোটবেলায় আমি একবার মেহমানী উৎসবে গিয়েছিলাম আজকালকার পার্থক পার্থিকারা মেহমানী শব্দটার সঙ্গে পরিচিত কি না জানি না কাজেই একটু ব্যাখ্যা করে নেই আগেকার আমলে বিত্তবান লোকদের একটা প্রবণতা ছিল তাদের বিত্তের বিষয় অন্যদের জানানো	1
রাজা বলল ভাল খবর মেম সাব আমি লটারী জিতেছি মেম সাব আমার টিকিটে ফার্স্ট প্রাইজ উঠেছে হ্যাঁ মেম সাব বিশ লাখ টাকা রাণী বিস্ময়ে চোখ বড় বড় করে বলল সত্যি রাজা বলল সত্যি রাণী বলল খুব ভাল কথা	1
পার্বতী কাছে আসিয়া বসিল আঁচলে যাহা বাঁধা ছিল তৎক্ষণাৎ দেবদাসের চক্ষে পড়িল কোন কথা জিজ্ঞাসা না করিয়া সে তাহা খুলিয়া থাইতে আরম্ভ করিয়া কহিল পারু পণ্ডিতমশাই কি বললে রে? জ্যেষ্ঠামশায়ের কাছে বলে দিয়েচে	0

Where,

0 = Sadhubhasha

1 = Cholitobhasha

Data Statistics:

Table 3.2: Data statistics

Number of Instance	Class
500	Sadhubhasha
501	Cholitobhasha

We used total 1000 paragraphs or instances for our dataset where half of them were in Sadhubhasha and rest was Cholitobhasha.

Data Pre-Processing

After data collection, we need to preprocess it again. We have removed the punctuation, brackets and stopwords so that while we train the model, so that we could find maximum accuracy. Finally, data preprocess was done in two-part; denoising and normalization.

Denoising

Denoising is a process by which we can remove any kind of html tags and brackets that could have gathered with dataset. It generally happens when we scrap data from different websites. The pseudo code for denoising are:

1. Import regular expression and string library
2. Import beautiful soup library
3. Define a function for beautiful soup
`soup = BeautifulSoup(text.strip(), "lxml")`
4. Define a function for brackets
`return re.sub("\[[^]]*\]", "", text`

Normalization

Data normalization is a process by which data attributes are organized in a data model or dataset. Data normalization increase the data consistency and reduce or eliminate data redundancy. Data normalization also helps to object-to-data mapping. For our dataset we used two function, one for removing punctuations and other for stop words.

Stop words are basically those words which are filtered out before or after NLP (natural language processing) data. Stop words are normally known as the most common words.

For our Bangla text classification, we have created a list of these stop words we have to eliminate. List are given below.

অতএব	একবার	করলেন	চলে	তাহলে	নাকি	বলে	যাবে	হন
অথচ	একে	করা	চান	তাহা	নাগাদ	বলেছে	যায়	হবে
অথবা	এক	করাই	চায়	তাহাতে	নানা	ন	যার	হবেন
অনুযা	এখন	করায়	চার	তাহার	নিজে	বলেন	যারা	হয়
য়ী	এখনও	করার	চালু	তিনিএ	নিজেই	বসে	যিনি	হয়তো
অনেক	এখানে	করি	চেয়ে	তিনি	নিজে	বহু	যে	হয়নি
অনে	এখানে	করিতে	চেপ্টা	তিনিও	দের	বা	যেখানে	হয়ে
কে	ই	করিয়া	ছাড়া	তুমি	নিজের	বাদে	যেতে	হয়েই
অনে	এটা	করিয়ে	ছাড়াও	তুলে	নিতে	বার	যেন	হয়েছি
কেই	এটাই	করে	ছিল	তেমন	নিয়ে	বি	যেমন	ল
অন্তত	এটি	করেই	ছিলেন	তো	নিয়ে	বিনা	র	হয়েছে
অন্য	এত	করেছি	জন	তোমার	নেই	বিভিন্ন	রকম	হয়েছে
অবাধি	এতটাই	লেন	জনকে	থাকবে	নেওয়া	বিশেষ	রয়েছে	ন
অবশ্য	এতে	করেছে	জনের	থাকবে	নেওয়া	বিষয়টি	রাখা	হল
অর্থাৎ	এদের	করেছে	জন্য	ন	র	বেশ	রেখে	হলে
আই	এব	ন	জনাও	থাকা	নেওয়া	বেশি	লক্ষ	হলেই
আগামী	এবং	করেন	জে	থাকায়	নয়	ব্যবহার	শুধু	হলেও
আগে	এবার	কাউকে	জানতে	থাকে	পক্ষে	ব্যাপা	শুরু	হলো
আগেই	এমন	কাছে	জানা	থাকেন	পরে	রে	সঙ্গে	হাজার
আছে	এমনকী	কাছে	জানা	থেকে	পরে	ভাবে	সঙ্গেও	হিসাবে
আজ	এমনি	কাজ	নো	থেকেই	পরেই	ভাবেই	সব	হৈলে

Fig 3.1: stop words for Bangla

আপনা	এল	কারণ	জানিয়ে	দিতে	পাওয়া	মধ্যভা	সম্প্রতি
র	এস	কি	ছে	দিন	পাচ	গে	সহ
আপনি	এসে	কিংবা	জে	দিয়ে	পারি	মধ্যে	সহিত
আবার	ঐ	কিছু	জনজন	দিয়েছে	পারে	মধ্যেই	সাধারণ
আমরা	ও	কিছুই	টি	দিয়েছে	পারেন	মধ্যেও	সামনে
আমা	ওঁদের	কিন্তু	ঠিক	ন	পি	মনে	সি
কে	ওঁর	কী	তখন	দিলেন	পেয়ে	মাত্র	সুতরাং
আমাদে	ওঁরা	কে	তত	দু	পেয়ে	মাধ্যমে	সে
র	ওই	কেউ	তথা	দুই	প্রতি	মোট	সেই
আমার	ওকে	কেউই	তবু	দুটি	প্রথম	মোটাই	সেখান
আমি	ওখানে	কেখা	তবে	দুটো	প্রভৃতি	মোটাই	সেখানে
আর	ওদের	কেন	তা	দেওয়া	প্রযুক্ত	যত	সেটা
আরও	ওর	কোটি	তাঁকে	দেওয়ার	প্রাথমি	যতটা	সেটাই
ই	ওরা	কোন	তাঁদের	দেওয়া	ক	যথেষ্ট	সেটাও
ইত্যাদি	কখনও	কোনও	তাঁর	দেখতে	প্রায়	যদি	সেটি
ইহা	কত	কোনো	তাঁরা	দেখা	প্রায়	যদিও	স্পষ্ট
উচিত	কবে	ক্ষেত্রে	তাঁাহা	দেখে	ফলে	যা	স্বয়ং
উত্তর	কমনে	কয়েক	রা	দেন	ফিরে	যাঁর	হইতে
উনি	কয়েক	খুব	তাই	দেয়	ফের	যাঁরা	হইবে
উপর	কয়েক	গিয়ে	তাও	দ্বারা	বক্তব্য	যাওয়া	হইয়া
উপরে	টি	গিয়েছে	তাকে	ধরা	বদলে	যাওয়ার	হওয়া
এ	করছে	গিয়ে	তাতে	ধরে	বন	যাওয়া	হওয়ায়
এঁদের	করছেন	গুলি	তাঁদের	ধামার	বরণ	যাকে	হওয়ার

Fig 3.2: stop words for Bangla

We kept these words to a text file which we took as an input when we normalized the data. The

pseudo code for normalization is,

1. import string and regular expression library
2. define a function for punctuations
3. sentence = re.sub(r'|!|",',sentence)

4. run an if statement and return sentence.translate(str.maketrans(",string.punctuation))
5. define a function for stop words
6. ('stopwords-bn.txt', 'r', encoding='utf8', errors='ignore') as f:
bn_stopwords = f.read().split()
7. Run a for loop and join words without stop words

3.4 Proposed Methodology

Methodology

Text classification can be done in few different ways. For automatic text classification there are three approach are widely recognized. They are namely Rule based, Machine learning based and Hybrid system. For our project we have choose machine learning approach. We first converted our data to vectors using Bag of Words. As Bag of Words has few drawbacks, then we have used IF-TDF for numeric representation of Bangla dataset. After that, we split our dataset into two-part, training and testing. For training and classifying we tested different classifier algorithms.

Bag of Words

BoW or known as Bag of Words is a way of extracting features from text for machine learning algorithms. A Bag of Words mainly a representation of text that illustrate the occurrence of words in a text document. It implicates two things, they are,

1. A vocabulary of known words.
2. A measure of the presence of known words.

The main reason it is called a “bag” of words, because any information about the structure or the order of words in the document is neglected. The model is only concerned about if the known words occur in the text document, not where in the text document.” A very common feature extraction procedures for sentences and documents is the bag-of-words approach (BOW). In this approach, we looked at the histogram of the words within the text, i.e. considering each word count as a feature” The bag-of-words model can be as simple or complex depending on the dataset and researcher. The complexity of BoW comes both in deciding how to design the vocabulary of known

words or tokens and how to score the presence of known word or tokens. Let's see an example,

রাজা বলল ভাল খবর মেমসাব আমি লটারী জিতেছি মেমসাব

let's take this line as an example. Now BoW will first collect all the vocabulary. For this line those vocabularies are,

রাজা, বলল, ভাল, খবর, মেমসাব, আমি, লটারী, জিতেছি

After collecting all vocabularies BoW will score all the words. Like,

Table 3.3: Vocabulary Scoring

Vocabulary	Score	vocabulary	Score
রাজা	1	মেমসাব	0
বলল	1	আমি	0
ভাল	1	লটারী	0
খবর	0	জিতেছি	1

So, the line we took its binary vector representation would be like

Table 3.4: Sentence binary vector representation

[1 1 1 0 0 0 0 1 0]

TF-IDF

TF-IDF also known as term frequency-inverse document frequency. TF-IDF weight is a statistical measure normally used to evaluate how important a word is to a text document in a dataset. The importance increases proportionally to the number of times a word appears in the text document but is cancel out by the frequency of the word in the dataset. The bag of words approach works well for converting text into numerical from but it also has a drawback. Which is It doesn't take into account the fact that the

word might also be having a high frequency of occurrence in other text documents. TF-IDF handle this issue by multiplying the term frequency of a word by the inverse document frequency.

The term frequency is calculated as:

$$\text{Term frequency} = \frac{(\text{number of Occurrence of a word})}{\text{Total words in a document}} \quad (1)$$

And the Inverse Document Frequency is calculated as:

$$\text{IDF (word)} = \frac{\text{Total number of documents}}{\text{Number of documents containing the words}} \quad (2)$$

Classification Algorithms

Random Forest

Random forest is a supervised machine learning algorithm mostly use for classification and regression. Random forest is an ensemble learning method algorithm means it use multiple learning algorithms to get a better predictive performance. Random forest works in a very easy way. From its name we can understand it creates forest means it generate many decision trees then merge them together to get a high accuracy performance.

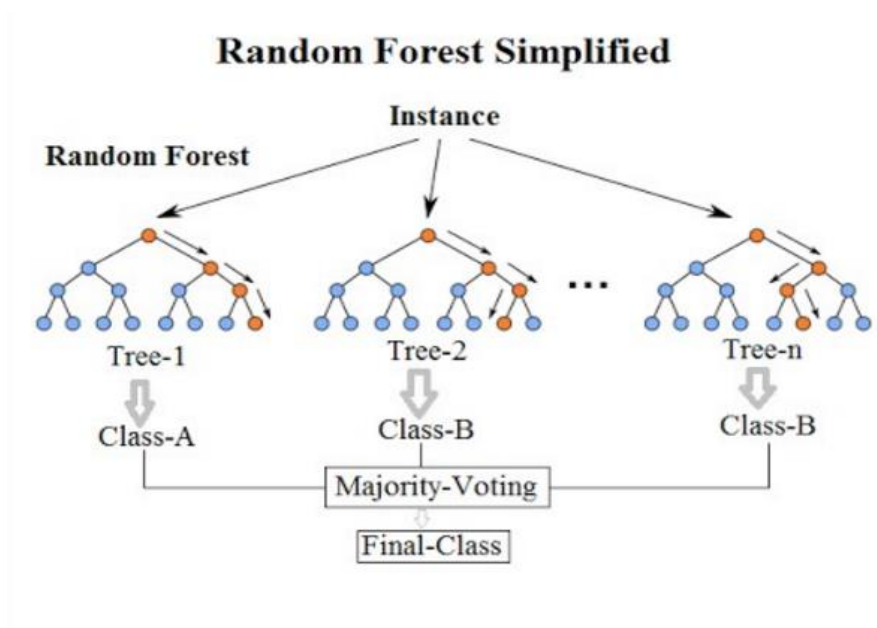


Fig 3.3: Random Forest algorithm

Random forest applies the general technique of bootstrap aggregating or bagging to teach the trees which trains the dataset. After finishing training predictions for unseen samples x' can be predict by averaging the predictions from all the individual regression trees on x' :

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x') \quad (3)$$

or by taking the majority vote in the case of classification trees. And for creating the standard deviation we need the summation of the uncertainty of prediction from all the individual regression trees on x' :

$$\sigma = \sqrt{\frac{\sum_{b=1}^B (f_b(x') - \hat{f})^2}{B-1}} \quad (4)$$

Where,

B= number of trees

f_b = regression tree

Naive Bayes classifier

Naive Bayes classifiers are a group of probabilistic supervised machine learning algorithms based on applying Bayes' theorem with naive assumption of conditional independence between every pair of features. The Bayes theorem:

$$P(Y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)} \quad (5)$$

where,

y= given class variable

x_1, \dots, x_n = dependent feature vector

For our dataset we used two Naive Bayes classifier Gaussian Naive Bayes and Multinomial Naive Bayes.

Gaussian Naive Bayes

When working with continuous data an assumption is made that the continuous values related to each class are distributed by following Gaussian distribution. The equation for it is:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (6)$$

Multinomial Naive Bayes

Multinomial naive Bayes algorithm implements the multinomial distribution method for distributing data. This algorithm is widely used for text classification. The equation is:

$$p(x / C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{ki}^{x_i} \quad (7)$$

Decision Tree

Decision tree classifier is a non-parametric predictive supervised machine learning algorithm which uses decision trees to predict. Decision trees predict a variable's value by learning decision rules extracted from data features. Decision trees are mainly used for classification and regression. For our work we only work with the classification part of the decision tree. The mathematical formulation for classification is:

$$p_{mk} = 1/N_m \sum_{x_i \in R_m} I(y_i - k) \quad (8)$$

Where,

m = node

R_m = region and

N_m = observation

Now, to measure impurity using Gini:

$$H(X_m) = \sum_k p_{mk} (1 - p_{mk}) \quad (9)$$

Entropy:

$$H(X_m) = -\sum_k p_{mk} \log(p_{mk}) \quad (10)$$

And Misclassification:

$$H(X_m) = 1 - \max(p_{mk}) \quad (11)$$

Where X_m represents the training data in node m .

Support Vector Machines (SVMs)

Support vector machines are a group of supervised machine learning algorithms which are widely used for classification and regression. Support vector machine algorithms are very efficient while working in high-dimensional space. Support vector machine

algorithms are also very memory efficient. For our text classification we used SVC method.

To classify Support vector machine creates a hyperplane in a high dimensional space which allow us to achieve more efficient and accurate classify result.

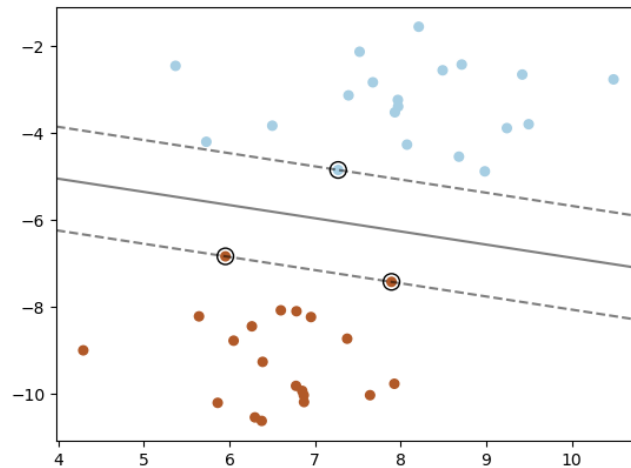


Fig 3.4: Support vector machine creating hyperplane to classify

k-nearest neighbors

K-nearest neighbors is a non-parametric supervised algorithm used for classification and regression. K nearest neighbors mainly work with similarity. This algorithm works on instance base. It does not create any model but save the instance of the training data. From the data it calculates highest vote of the nearest neighbors of each point then a query point is assigned the data class which has the most representatives within the nearest neighbors of the point k-nearest neighbors calculated by a distance function. If $K = 1$, then the case is simply assigned to the class of its nearest neighbor. The functions are:

$$\text{Euclidean} = \sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (12)$$

$$\text{Manhattan} = \sum_{i=1}^k |x_i - y_i| \quad (13)$$

$$\text{Minkowski} = (\sum_{i=1}^k (|x_i - y_i|^q)^{1/q}) \quad (14)$$

All these functions work well with continuous variable. For categorical variable we need to use Hamming distance. The equation is:

$$D_H = \sum_{i=1}^k |x_i - y_i| \quad (15)$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

The best value is determined by the large value of K which reduce the overall noise.

Evaluation Metrics

Confusion matrix

A confusion matrix is a table by which we describe the performance of a dataset which is used for classification. There are four elements in confusion matrix. TP or True Positive, TN or True negative, FP or False Positive and FN Or False negative.

Abbreviation	Name	Description
TP	True Positives	Number of correct classifications predicted as positive (or Yes)
TN	True Negatives	Number of correct classifications predicted as negative (or No)
FP	False Positive	Number of examples that are incorrectly predicted as positive when it is actually negative
FN	False Negative	Number of examples that are incorrectly predicted as negative when it is actually positive

Fig 3.5: Details of confusion matrix

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

Now False positives and false negatives, these values appear when actual class contradicts with the predicted class.

False Positives (FP) – When actual class is no and predicted class is yes.

False Negatives (FN) – When actual class is yes but predicted class in no.

Accept this there are few more term that we need to understand before we discuss about. They are,

Accuracy

Accuracy is the most direct approach for performance measure which is a ratio of correctly predicted class to the total class. If we have high accuracy then our model is best. But accuracy is a great measure only when we have symmetric datasets where values of false positive and false negatives are almost same. That's why, we have to look at other parameters to evaluate the performance of our model.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (16)$$

Precision

Precision is the ratio of correctly predicted positive instances to the total predicted positive instances. The question that this metric answer is of all instance that labeled as Sadhubhasha and Cholitobhasha, how many of them are actually Sadhubhasha and Cholitobhasha instance? High precision means low false positive rate.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (17)$$

Recall

Recall also known as Sensitivity. Recall is the ratio of correctly predicted positive instances to the all instances in actual classes. The question recall answers is: Of all the instances of Sadhubhasha and Cholitobhasha that truly classified, how many did we label?

$$\text{Recall} = \frac{TP}{TP+FN} \quad (18)$$

F1 score

F1 Score is basically the average of Precision and Recall. That means, this score takes both false positives and false negatives into count. basically, it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy.

$$F1 \text{ score} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \quad (19)$$

now we understand the terms let's see the result we got from our classifying model we used for classifying Sadhubhasha and Cholitobhasha.

3.5 Statistical analysis

In our dataset we took 1001 instance as data. 500 of them are Sadhubhasha and 501 are Cholitobhasha. These 1000 instance combinedly have over 100000 thousand words of Bangla language. We took 80% means almost 80000 word for our training and rest 20000 words data are kept for testing.

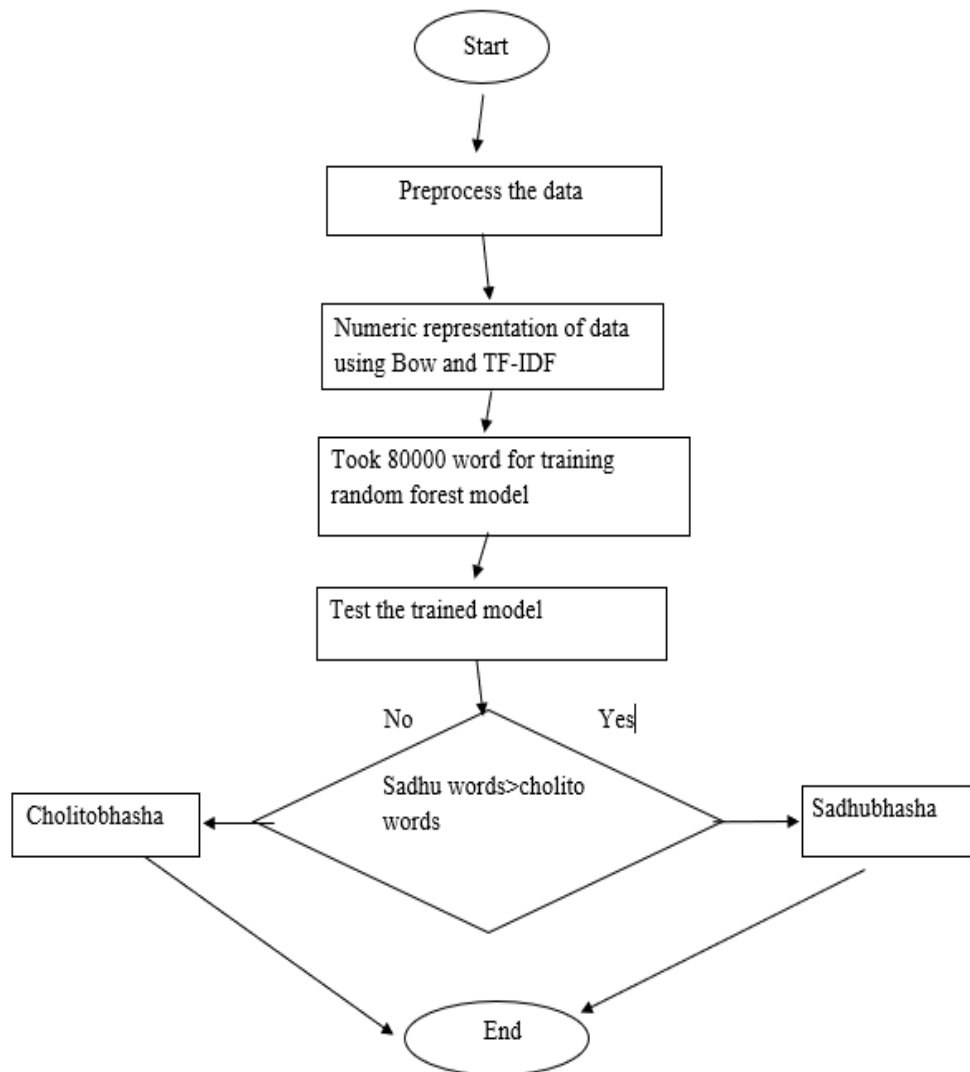


Fig 3.6: Flow Chart of Proposed Model

3.6 Implementation Requirements

After the reviewing all the all necessary statistical or theoretical concepts and methods, we created a list of Hardware, Software and developing tools we need for classifying Sadhubhasha and Cholitobhasha. The probable necessary things are:

Hardware/Software Requirements

- Operating System (Windows 7 or above)
- Ram (more than 4 GB)
- Web Browser (preferably chrome)

Developing Tools

- python 3.7
- Anaconda
- Jupiter notebook
- NLTK
- Pandas
- NumPy

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

In Chapter four we will discuss about descriptive analysis of our project. We will state about our experimental result and finally we will close the chapter with a summarization of result.

4.2 Experimental Results

To measure the effectiveness and accuracy of the algorithms we choose few methods such as precision, recall, f1 score and support. This method will help us to understand which algorithms can classify bangla text more accurately.

Random forest classifier:

Random forest classifier is a supervised machine learning algorithm that applies the bootstrap aggregating or bagging methods to train data. The accuracy we achieve from random forest is 98.5%. The detailed result:

Confusion matrix = $\begin{bmatrix} 97 & 3 \\ 0 & 101 \end{bmatrix}$

Here,

0 = Sadhubhasha

1 = Cholitobhasha

Table 4.1: result achieved from random forest algorithm

Metrics	0	1
Precision	1.00	0.97
Recall	0.97	1.00
F1 score	0.98	0.99
Support	100	101

Naive Bayes

We implemented two classifier algorithms from naive Bayes classifier. Gaussian Naive Bayes and multinomial Naive Bayes.

Multinomial Naive Bayes

For multinomial Naive Bayes we achieve accuracy of 99.5%.

Confusion matrix = $\begin{bmatrix} 99 & 1 \\ 0 & 101 \end{bmatrix}$

Table 4.2: result achieved from Multinomial Naive Bayes algorithm

Metrics	0	1
Precision	1.00	0.99
Recall	0.99	1.00
F1 score	0.99	1.00
Support	100	101

Gaussian Naive Bayes

For Gaussian naive Bayes we achieve accuracy of 91.04%.

Confusion matrix = $\begin{bmatrix} 88 & 12 \\ 6 & 95 \end{bmatrix}$

Table 4.3: result achieved from Gaussian Naive Bayes algorithm

Metrics	0	1
Precision	0.94	0.89
Recall	0.88	0.94
F1 score	0.91	0.91
Support	100	101

Support Vector Machine

We use the SVC or support vector classifier method from Support Vector Machine algorithm. It achieves 97.014% accuracy.

Confusion matrix = [[94 6]
[0 101]]

Table 4.4: result achieved from Support Vector Machine algorithm

Metrics	0	1
Precision	0.94	0.89
Recall	0.88	0.94
F1 score	0.91	0.91
Support	100	101

K-nearest neighbor

For K-nearest neighbor algorithm we achieve accuracy of 94.52% .

Confusion matrix = [[96 4]
[7 94]]

Table 4.5: result achieved from K-nearest Neighbor algorithm

Metrics	0	1
Precision	0.93	0.96
Recall	0.96	0.93
F1 score	0.95	0.94
Support	100	101

Decision Tree

For decision tree algorithm we achieve accuracy of 91.54% .

Confusion matrix = [[91 9]
[8 93]]

Table 4.6: result achieved from decision tree algorithm

Metrics	0	1
Precision	0.92	0.91
Recall	0.91	0.92
F1 score	0.91	0.92
Support	100	101

From the result tables we can find out that all algorithms achieve good accuracy. For our bangla data Multinomial Naive Bayes and Random forest algorithms achieves the highest accuracy 99.5% and 98.5% respectively. Support vector Machine and K-nearest neighbor algorithms performed well with an accuracy of 97% and 94% respectively. Decision tree performed good but had the lowest accuracy between all the algorithms.

4.3 Descriptive Analysis

In this section we will further discuss about the result of our proposed method. We will also focus the word cloud analysis about project and finally a graphical representation of most used words in our dataset.

For our bangla data Multinomial Naive Bayes and Random forest algorithms achieves the highest accuracy 99.5% and 98.5% respectively. Support vector Machine and K-nearest neighbor algorithms performed well with an accuracy of 97% and 94% respectively. Decision tree performed good but had the lowest accuracy between all the algorithms.

Word cloud Analysis

This is a process by which we could find the most used word for Sadhubhasha and Cholitobhasha separately. It is a simple process by which we can easily determine those words that helped us to classify Sadhubhasha and Cholitobhasha. Now let's have a look to most used 20 words in Sadhubhasha and Cholitobhasha from our dataset.


```
Out[25]: ['কারিয়া',  
'বলিল',  
'হইয়া',  
'বলিলেন',  
'কথা',  
'দিয়া',  
'বলিয়া',  
'হইল',  
'ত',  
'একটা',  
'করিল',  
'আসিয়া',  
'ললিতা',  
'শেখর',  
'লাগিল',  
'কহিল',  
'লইয়া',  
'এক',  
'মা',  
'করিলেন']
```

Fig 4.1: List of most used words in Sadhubhasha

```
Out[28]: ['হয়ে',  
'একটা',  
'কথা',  
'দিয়ে',  
'এক',  
'ভাল',  
'বড়',  
'হয়েছে',  
'মত',  
'নবুমামা',  
'সময়',  
'সফিক',  
'মা',  
'জন্যে',  
'যায়',  
'টাকা',  
'সাহেব',  
'দেখি',  
'মামা',  
'ঘরে']
```

Fig 4.2: List of most used words in Cholitobhasha

Now if we plot these words in to a image by the size of every word we will easily find out which word is repeated or used most in Sadhubhasha and Cholitobhasha from our dataset

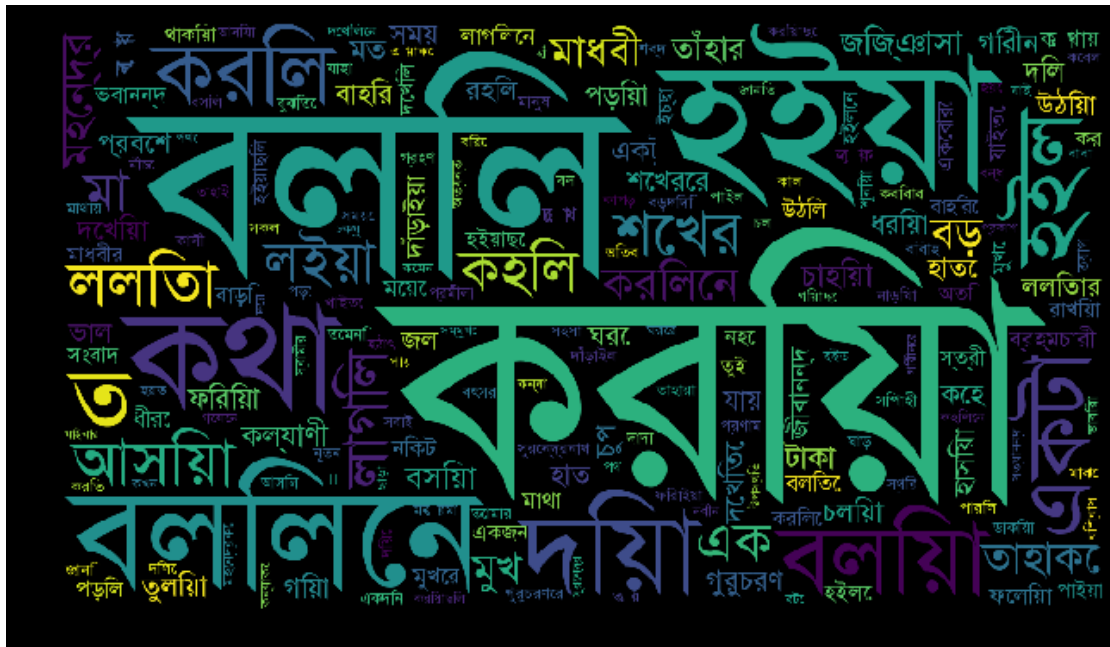


Fig 4.3: image plot of most used words in Sadhubhasha



Fig 4.4: Image plot of most used words in Cholitobhasha

4.4 Summary

After getting the accuracy by using the Bag of Words and Term Frequency-Invers Document Frequency for vector representation and from different classifier algorithm, we trained our data and classify it between Sadhubhasha and Cholitobhasha in our research dataset from different novels of Bangla language. We achieved from 91% to 99.5% accuracy for different algorithms. But if we want to increase the accuracy, we need to prepare the dataset more accurately. Also there is a need to increase the size of dataset for statistical significance. More preprocessed data will also help to achieve higher accuracy to classify Sadhubhasha and Cholitobhasha.

CHAPTER 5

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH

5.1 Summary of the Study

Our main target was to build a model which will help to classify two form of Bangla language, Sadhubhasha and Cholitobhasha. We took the machine learning approach to classify the two form Sadhubhasha and Cholitobhasha. In this approach we vectorize the Bangla data then we split the data for training and testing. Following that, we implemented machine learning algorithm like Random forest, Naive Bayes, Support Vector Machine, Decision Tree and K nearest neighbor. Each of these algorithms performed well. Random Forest achieves 98.5% accuracy, Multinomial Naive Bayes achieves 99.5% accuracy, Support Vector Machine achieves 97% accuracy, K nearest neighbor achieves 94.5% accuracy and Decision Tree achieves 91% accuracy.

5.2 Conclusions

The accuracy level that we have achieved using these classifier algorithms is significant. It is indicative that our proposed method performed better than other methods used for bangla text classification. After completing the research, we have learned a lot of things about Natural Language processing and Machine Learning. Now we can preprocess the data and also train a model for classifying text documents. We hope this will also help in further research in Bangla language and in text classification area.

5.3 Recommendations

Few recommendations for Bangla text classification are:

1. Create a large dataset for high accuracy
2. Try to remove all other language word written in Bangla for better accuracy
3. Find and list all the stop words this will also help you to increase the accuracy

5.4 Implication for Further Study

Few implications that possible in further studies are:

1. Adding more categories like combined Sadhubhasha and Cholitobhasha data in this project, can make this more efficient.
2. Using more classifier algorithms on this dataset, can get a better understanding on which classifier perform well and give us the best and higher accuracy.

REFERENCES

- [1].Abu Nowshed Chy, Md. Hanif Seddiqui, Sowmitra Das, " Bangla News Classification using Naive Bayes classifier," 16th Int'l Conf. Computer and Information Technology, 8-10 March 2014, Khulna, Bangladesh.
- [2].Andrew McCallum and Kamal Nigam," A Comparison of Event Models for Naive Bayes Text Classification," Published 1998.
- [3].Andronicus A. Akinyelu and Aderemi O. Adewumi, " Classification of Phishing Email Using Random Forest Machine Learning Technique," Journal of Applied Mathematics Volume 2014, Article ID 425731, 6 pages.
- [4].Baoxun Xu, Xiufeng Guo, Yunming Ye and Jiefeng Cheng " An Improved Random Forest Classifier for Text Categorization," journal of computers, vol. 7, no. 12, December 2012.
- [5].Suresh Merugu, M. Chandra Shekhar Reddy, Ekansh Goyal and Lakshay Piplani, "Text Message Classification Using Supervised Machine Learning Algorithms," International Conference on Communications and Cyber Physical Engineering 2018, ICCCE 2018: ICCCE 2018 pp 141-150.
- [6].M. Ikonomakis, S. Kotsiantis and V. Tampaka , " Text Classification Using Machine Learning Techniques," WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005, pp. 966-974.
- [7].Timothy P. Jurka, Loren Collingwood, Amber E. Boydston, Emiliano Grossman, and Wouter van Atteveldt," RTextTools: A Supervised Learning Package for Text Classificatio" The R Journal Vol. 5/1, June ISSN 2073-4859.

Plagiarism Report

Shadhu chalito

ORIGINALITY REPORT

19%	18%	14%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	blog.exsilio.com Internet Source	5%
2	computer-trading.com Internet Source	3%
3	stackabuse.com Internet Source	2%
4	www.academypublisher.com Internet Source	2%
5	oysauce.com Internet Source	2%
6	medium.com Internet Source	1%
7	singaporetranslation.com Internet Source	1%