Computer Science and Engineering   Undergraduate Project Report

2018-05

Bangla News Classification Using Machine Learning

Ahmad, Mostak

Daffodil International University

BANGLA NEWS CLASSIFICATION USING MACHINE LEARNING

BY

Mostak Ahmad

ID: 142-15-3800

Fayjun Nahar Mishu

ID: 142-15-3665

S. M. Shakib Limon

ID: 142-15-3842

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

Ahmed Al Marouf

Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2018

APPROVAL

This Project titled **"Bangla News Classification Using Machine Learning"**, submitted by Mostak Ahmad, Fayjun Nahar Mishu and S. M. Shakib Limon to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering (BSc) and approved as to its style and contents.

BOARD OF EXAMINERS

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

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

Md. Zahid Hasan   Internal Examiner
Assistant Professor
Department of Computer Science & Engineering
Faculty of Science & Information Technology
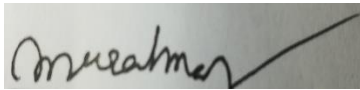Daffodil International University

Dr. Mohammad Shorif Uddin   External Examiner
Professor
Department of Computer Science & Engineering
Jahangirnagar University
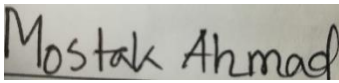
ii

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.
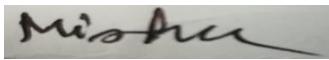
Supervised by:

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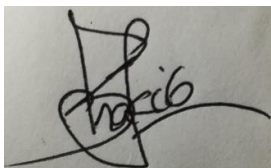Department of CSE
Daffodil International University

Submitted by:

Mostak Ahmad
ID: 142-15-3800
Department of CSE
Daffodil International University

Fayjun Nahar Mishu
ID: 142-15-3665
Department of CSE
Daffodil International University

S. M. Shakib Limon
ID: 142-15-3842
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 and keen interest of our supervisor in the active learning model design influenced 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 **Dr. Syed Akhter Hossain, Head, Department of CSE**, for his kind help to finish our project and also to other faculty members and the staffs 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

What is the distance between countries to countries from south to north, east to west in the earth? If you find the answer of this question according to the perspective of the present time, you will see that, actually, there is no distance at all. At present, people get the all sorts of news happening around the world instantly within couple of seconds. And it has become possible only because of virtual news portal. That is true that the online news portals are publishing news on live, but it is disappointing that users do not like all sorts of news published in the news portal. At that time, it has become need to a platform that can easily identify the user's choice on news and publish only according to their choice. To classify the news by the user's choice needs to analysis the news text.

Lots of works has been done in English news by this time but there have very limited works on Bangla news. These make us inspired to do a research project on this topic though Bangla is one of the 8 major spoken languages around the world. In our research project, we deal with Bangla news collected from Prothom Alo newspaper. From preprocessing the news text, we try to do all sorts of procedures to classify the news text using Machine Learning classifier, "Naive Bayes classifier". Finally, we develop a user interface to take the news text and show the class of that news.

TABLE OF CONTENTS

CONTENTS   **PAGE**

CHAPTER

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Today, we are living in such a universe where there are no limits among the nations. Maybe, an occurrence may have been happened thousands miles far away however the truth of the present world is-it takes not as much as seconds to spread this news all through the world. We can get or read a large number of news from wherever on the planet with the favors of web, PC and in addition with the cutting edge innovation. For this situation, News Portals are mindful to spread the news quickly through the web. Bangla news entryways are not falling behind. In nowadays, there are heaps of Bangla news entryway exist in the social or virtual life. These news gateways are constantly mindful to the new events happened of our environment. As indicated by our squire's point of view, these news entry are extremely such a great amount of caution to distribute moment hot and selective news to their comparing entryway. Some Bangladeshi news portals are:

- Daily Prothom Alo.
- Ittefaq.
- Samakal.
- Dainik Amader Shomoy.
- Daily Naya Diganta.
- Jai Jai Din.
- Bangladesh Pratidin etc.

These news entries are persistently distributing the state-of-the-art news. Our venture is managing this Bangla news. We built up an exceptionally basic site that can distinguish or recognize the class of news that has given by a client. Before building up this site, we considered on its hypothetical ideas. By concentrate different sorts of papers identified with this work, we made the techniques utilizing machine learning ways to deal with arrange the news articles.

1.2 Objectives:

- To study how to classify or categorize Bangla news using some classifier algorithm.
- To develop a platform that will be able to detect the category of given Bangla news.
- To visualize some analytical analysis of Bangla News classification classified by classifier algorithms.

1.3 Motivation

We see that the news entries distribute a wide range of news. That implies, all out news is being distributed in these news entryways. Be that as it may, all individuals don't incline toward a wide range of news. A few people want to peruse sports news most than political news. A few people get a kick out of the chance to peruse political news than alternate news. A few people get a kick out of the chance to peruse amusement news. It really relies upon one's decision. In any case, now and then, it has turned out to be such a great amount of exhausting to see the news that, really, isn't favored by the client. The news entry turns into the most proficient in the event that it demonstrates the news as indicated by the particular client's decision. In any case, for this, the main errand is to distinguish the news class. We discover bunches of undertakings on news order in English. In any case, there is extremely poor work on Bangla. In the event that Bangla news order gets some exploration chips away at it, it can be utilized as a part of such of numerous genuine applications.

Other than this, we see that the present world is such a great amount of concentrating on proposal framework. Clients expect everything that the better things will be prescribed to them by the framework. To make a framework to be proposal skilled must be able to take choice without anyone else's input. To take choice independent from anyone else must need the information mining ability.

These made us intrigued to do such sort of research based work. Our work is completely related with machine learning methods and has a few information mining strategies as well.

1.4 Rationale of the Study

It is no doubt there are lots of works on Natural Language Processing (NLP) in English and these approaches or the processes are being used in many automated system as well as robotics system. But, Natural Language Processing on Bangla is very rare. To develop more automated application or make much more efficient of Machine Learning approaches in Bangla, there has no alternative to work with Bangla text. This made us to be interested to work with this Bangla News classification.

In the present time, we see that the notepad editors are much more intellectual. These have some features like auto corrections, grammar checking, auto suggestion etc. These features are the outcome of the blessing of Natural Language Processing. But these features are mostly seen for English. Such kind of features is very rare for Bangla text. These, actually, take us to work with Bangla news as well as Text.

1.5 Research Question
- ➢ Can we collect row data of Bangla News?
- ➢ Can we pre-process the row data to be used for the Machine Learning approaches?
- ➢ Can Multinomial Naïve Bayes Classifier algorithm be used on the pre-processed data?
- ➢ Can the Machine Learning process correctly detect or identify the category of the given Bangla dataset?

1.6 Expected Output

Expected result of this exploration based undertaking is to construct a calculation or making a total productive strategy that will order given Bangla news as for the assembled model of prepared dataset.

1.7 Report Layout

The report will be followed as follows

Chapter 1 provides the summary of this research based project. Introductory discussion is the key term of this first chapter. Apart from, what motivated us to do such a research based project is explained well in this chapter to. The most important part of this chapter is the Rationale of the Study. Then, what are the research questions and what is the expected outcome is discussed in the last section of this chapter.

Chapter 2 covers the discussion on what already done in this domain before. Then the later section of this second chapter shows the scope arisen from their limitation of this field. And very last, the root obstacles or challenges of this research are explained.

Chapter 3 is nothing but the theoretical discussion on this research work. To discuss the theoretical part of the research, this chapter elaborates the statistical methods of this work. Besides, this chapter shows the procedural approaches of the Machine Learning classifier-Multinomial Naïve Bayes. And in the last section of this chapter, to validate the model as well as to show the accuracy label of the classifier, confusion matrix analysis is being presented.

Chapter 4 is related with the outcome of the whole research and the project. Some experimental pictures are presents in this chapter to make realize the project.

Chapter 5 is based on conclusion topics of the project. This chapter is responsible to show the whole project report adhering to recommendation. The chapter is closed by showing the limitations of our works that can be the future scope of others who want to work in this field.

CHAPTER 2
BACKGROUND

2.1 Introduction

This chapter mirrors the related works that effectively done by a few specialists in the past time in this field. Also, giving an unmistakable clarification of this, this part will indicate what the impediments of these works were and in conclusion, this section depicts extent of our exploration and also its difficulties.

2.2 Related Works

It is the matter of sorrow that very few works on this field has accomplished by this time though in the present time, working on this field is increasing day by day. There are enough resources for English language [5] as there has been done many works in this field for English.

Recently, not only in Bangla language, but also in other language as like Chinese[17] ,Indonesian[7,8], Hindi[4,9], Urdu[10], Arabic[3] English-Hindi[6] and so on, are being included on Natural Language Processing related works. There are being enriched with resources day by day after doing more research works on this field.

Some related works relates to our research work are given below with a short description.

Analysis of N-Gram based text categorization for Bangla in a newspaper corpus

The desire objective for any order is only building an arrangement of models by utilizing preprocessed datasets. This datasets are the core of such of undertaking. At that point, the datasets are being partitioned into two sections preparing dataset and testing dataset. At last, these two sub datasets are utilized to manufacture the model. The intension of building such sort of model is to anticipate the class of various s. An exploration group from BRAC University chipped away at such a point. They essentially works in view of N-Gram based order [1].

Text categorization is considered as the grouping. It implies that it is in charge of consequently appointing into some predefined classifications or grouping as for the given sections. The

objective of arrangement alludes to this naturally characterize reports into the classes that are predefined and these procedure is being done based on their substance.

In their proposal, there principle center was that examination if n-gram based classification can be connected on Bangla. In addition, they likewise break down the execution of their work.

What is an N-gram?

When something is done on N-gram, first inquiry naturally raised that what, really, alludes to the N-gram. In the event that in no time saying, N-gram is only the sub-grouping. It is the sub-grouping of n-things in any given arrangement. There are some application in view of N-gram idea on computational phonetics and these models are utilized for anticipating words or foreseeing characters for the objective of different application. If I show an example in favor of my words, then I say, the word বাংলা would be composed of following character level n-grams.

Table 2.2.1: Different n-grams for the word" " (spaces are shown with"_").

Unigrams	ব , , , ল , , _
Bi-grams	_ব , , , ল , , _
Tri-grams	— , , ল , , _
Quad-grams	— , ল , , _

Thus, we can compress our idea on n-gram that it is the system of character succession of length n removed from a report. For this situation, characterizing the estimation of N is so much concerned. The estimation of n is reliant on specific corpus of records. To produce the n-gram vector for a report, first need a window of character long that is traveled through the text . The, it ought to slide forward by a settled number of character.

Why N-gram Based Text Categorization?

It is usually appeared to us that human languages have some words that occur much more frequently than others. To get a better understanding on this concept, Zipf's Law can be an example. It is actually indicates some common ways to express this idea. It can be re-state as follows: "The nth most common word in a human language text occurs with a frequency inversely proportional to n."

It can be said so, if f is considered as the frequency of the word and r is the rank of the word in the list ordered by the frequency, Zipf's Law states,

$$f = k / r$$

The implication of this law is that there remains always a group of words in a text and it is commonly seen that these words are dominates most of the other words of the language in terms of frequency of use.

Test Data

For their experiment, they firstly selected 25 test documents randomly. These 25 documents are from each of the six categories. These are defined from the 1 year Prothom Alo news corpus. Thus, the total numbers of test cases were 150. List of predefined categories and their content source are following.

Defined category	Prothom-Alo Editorials
Technology News Category	কম্পিউটার প্রতিদিন, প্রজন্ম ডট কম
Sports News Category	খেলা
Deshi News Category	বিশাল বাংলা
International News Category	সারা বিশ্ব
Entertainment News Category	বিনোদন
Business News Category	অর্থ ও বাণিজ্য

Figure 2.2.1: List of predefined categories.

Procedures of their work flow is as follow:
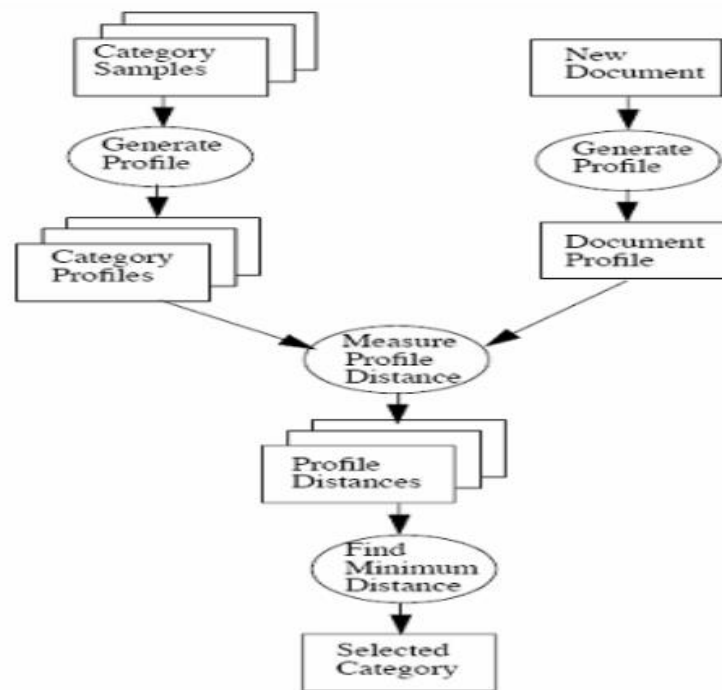
Figure 2.2.2: Classification procedure N-gram.

Observation:

In their experiment, they found that character level trigram perform better than any other n-grams. They thought the reason behind being better of trigram is it could hold more information for modeling the language.

A machine learning approach for authorship attribution for Bengali blogs

This examination work clarifies the portrayal of initiation attribution framework for Bengali blog writings. There, they had exhibited another Bengali blog corpus. This corpus contains just about 3000 entries. These 3000 entries were composed by three creators. In their examination, they have offered a framework that was with respect to arrangement framework. Their methodologies depended on lexical highlights. Lexical highlights alludes to character bigrams and trigrams, word n-grams and stop words. They accomplished over 99% precise outcomes on their dataset utilizing Multi layered Perceptrons (MLP) among the for classifiers [2].

They concluded by declaring that MLP can produce very good results for big data sets. They also claimed that lexical n-gram based features can be the best for any authorship attribution system.

2.3 Research Summary

The above discussion done on various types of research works from different research teams, it is being appeared to us that recently, research work on Bangla text is increasing day by day. Some good outcomes already prove this statement well. Though, enough resources are not present, but hope is that this field is becoming more resourceful each after passing a single day.

2.4 Challenges

The main challenges of this work are dealing with the datasets. To clean the dataset, we need some efficient approaches to perform it but there are not enough recognized approaches to do it. Another challenge of this work is not having enough resources regarding this topic.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter mainly deals with the theoretical knowledge of the research work. It will give the clear understanding of the concept of work. To make it more clear, very first, Research Subject and Instrumentation is explained shortly. Then we know that in the data mining or machine learning process data are the heart. For this reason, data collection process is described in this section. The chapter is being closed by giving the explanation of our project's statistical theories and besides, giving the clear concept of the implementation requirements.

3.2 Research Subject and Instrumentation

We mean by research subject is that research area that is being studied and researched for clear understandings. Not only for clear understanding, but also research subject is responsible for giving the right knowledge of various research parameters. On the other hand, Instrumentation refers to the required instruments or tools that are used by the researchers.

3.3 Data Collection Procedure

To look into on particular field, the quick and first thing is the Data. Information is, really, considered as the core of the machine learning process. What's more, for our examination, there has no option of information. In this way, it has turned into our most difficult errand for our examination. We gather our information from the most popular Bangla news entrance of Bangladesh named Prothom Alo. Our Bangla news is gathered from this site by utilizing corpus. We gathered just about 4 year's news from them. What's more, the news are put away as content record arrange.

3.4 Data Pre Processing

When we manage the column information, the achievement generally relies upon the pre-prepared information. The all the more productively information will be pre-prepared; the result will be more exact. In single word, it is the starting test for such sort of research based work. Our

column information has some html label name. So it must be expelled from the archive. This was our first mindful to expel the all html label name from the news. At that point, we need to keep up some to clean the superfluous space from the report. At that point it expels all new line to orchestrate it into a line. That implies, in the wake of aggregating any news document, each line will be dealt with as a news. At that point, in conclusion, for each individual news, this dole out a number for recognizing classification. We use (0-8) for six news categories. These are: 0→Politics, 1→ Crime, 2→ Sports, 3→ Entertainment, 4→Business, 5→ Life Style, 6→Accident, 7→National, 8→International.

Finally, subsequent to doling out a particular number for every news, this create tsv record arranged document that is tab isolated. This tsv record is, really, our pre-handled information with its class. Accordingly, every one of the six unmitigated news are being relegated particular number. The six news class brings about six tsv record. At that point we utilize another python record named join.py to join every one of the six tsv document into just a tsv document. For this, each of the six ordered news tsv records are stayed into a document. At that point, simply needs to say the name of the organizer, it marge the all documents augmentation as tsv into one record.

3.5 Work Flow of Identifying News Category

Removing Excluded Word

We have made a list that contains Bangla words that are actually meaningless with respect to identify a news category. We named after those words as Excluded words. We stored our all selected excluded words in a txt file named excluded_word_list_out.txt. When the program is being run, at first, remove all of the excluded words from our input file.

Figure 3.5.1: Show the excluded Bangla word.

Split and Join:

To remove the excluded words from the dataset, firstly, the whole dataset is split. This process spit the whole news into words. So, after splitting process, the whole news be the collection of

only words. Then, every word is checking according to excluded word list. If any word from the dataset is being matched with the excluded word list, then, this word is being removed from the dataset. After checking all words remaining in the dataset, joining process starts. The joining process is very simple- just join the words into each news.

Features Extraction:

This phase is the main part of the news classification. Mainly, this phase decides in which way classify will be done. We use the word count as our feature extraction. There are built in method for this in the sklearn. We just use import this method to use this method for our feature extraction.

Building Model:

After successfully feature extraction, we are ready for building our model. And this is being accomplished by training our machine. We split our dataset into 3:1. The three portion of our data set are used for our training dataset and the rest portion is for testing. That means, 75% data from the datasets are used training and rest 25% is considered as the testing.

Classifier Fitting:

In this stage, our machine is ready or fit for the classifier. We use several classifier such as Naïve Byes, Decision Tree, K-Nearest Neighbors, Support Vector Machine and Random Forest for classify our news text. Sklearn has built in classifier of this. We just import it and fit it.

Predict the Category

This is the final stage of our news classification approach. In this stage, our model is being prepared for testing Bangla text input data. According to the given input text, this model can classify this text using several classifier such as Naïve Byes, Decision Tree, K-Nearest Neighbors, Support Vector Machine and Random Forest.

Flow Chart:

```
                          ( Start )
                             |
                             v
                     / Take Input /
                             |
                             v
                 [ Preprocess the text file ]
                             |
                             v
                    [ Split the text ]
                             |
                             v
   +------------->  [ Read Excluded File ]
   |                         |
   |                         v
[ Remove the word ] <--Yes-- < Match >
                             |
                             No
                             |
                             v
                    [ Join the words ]
                             |
              +------------->|
              |              v
              No         < Is classify? >
                             |
                            Yes
                             |
                             v
                    [ Show the class ]
                             |
                             v
                          ( End )
```
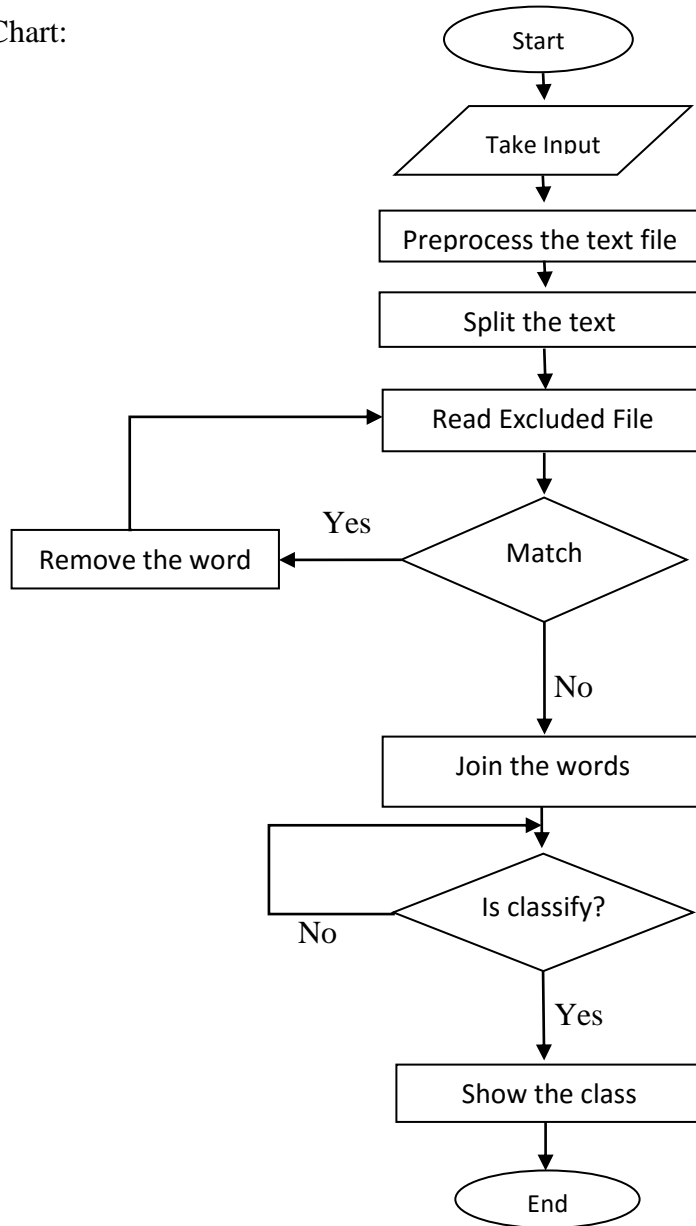
Figure 3.5.2: Proposed Working Flow chart for classification.
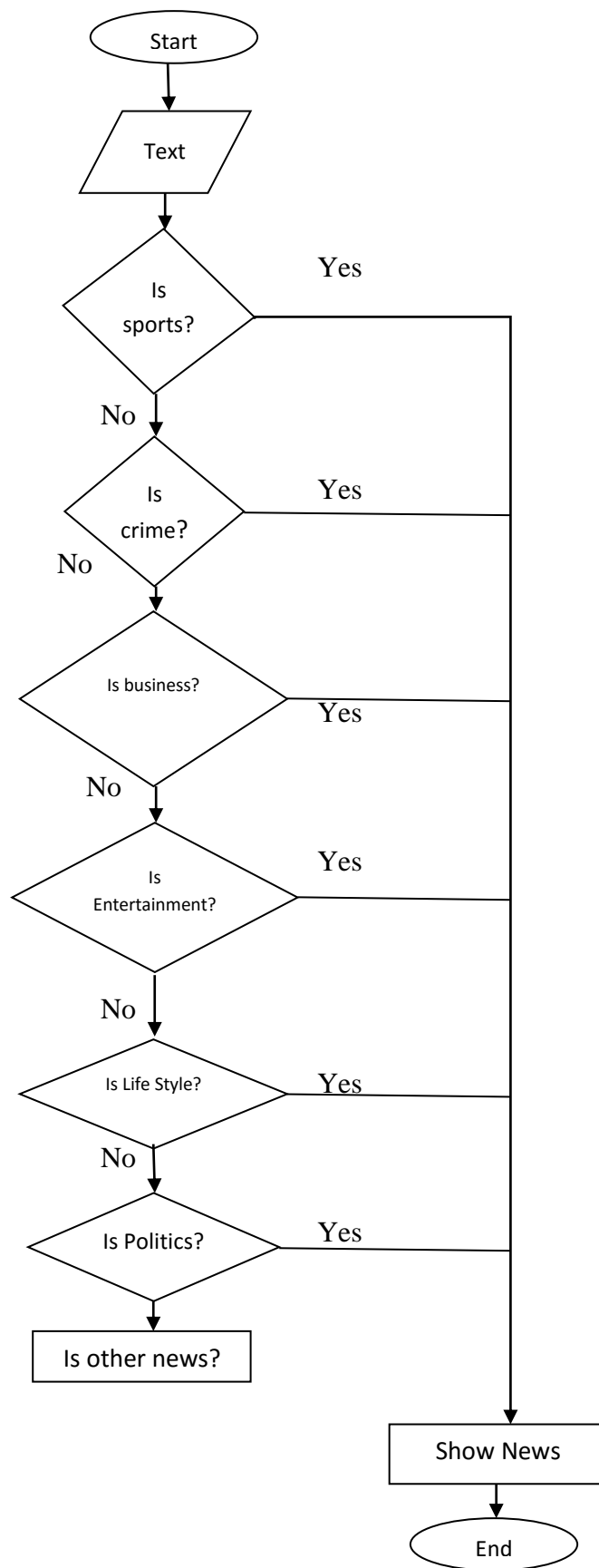
Figure 3.5.3: Classification process flowchart.

3.6 Implementation Requirements

After the proper analysis on all necessary statistical or theoretical concepts and methods, a list of requirement has been generated that must be required for such a work of Bangla News Classification. The probable necessary things are:

Hardware/Software Requirements

- ✓ Operating System (Windows 7 or above)
- ✓ Hard Disk (minimum 4 GB)
- ✓ Ram(more than 1 GB)
- ✓ Web Browser(preferably chrome)

Developing Tools

- ✓ Python Environment
- ✓ Spyder (Anaconda3)
- ✓ Django 1.11 (For UI)
- ✓ Notepad++
- ✓ Bootstrap

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

This chapter 4 mainly focuses on the descriptive analysis of the data used in the research as well as the experimental results of our project.

4.2 Raw Data

Our raw data are from the most renowned news portal of Bangladesh named Prothom Alo. We collect our data by using Corpus. After collecting data, news is stored on text document file. In these file, data are present with some html tag name. Our row data looks like:
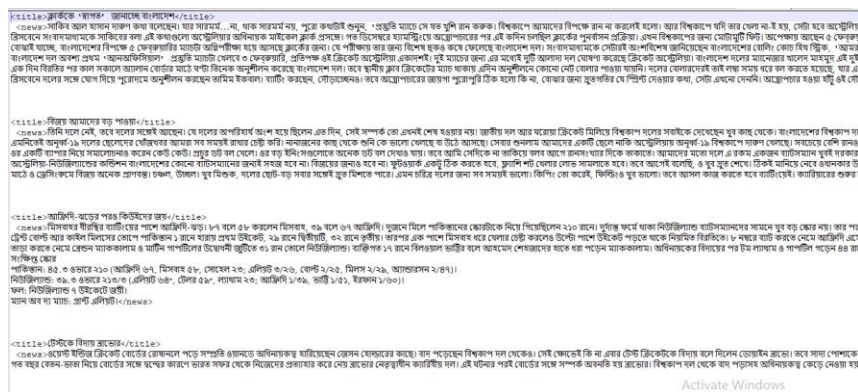
Figure 4.2.1: Experimental raw data.

So it has become obvious to clean the data. That means pre-processed the row data for preparing for the model.

4.3 Cleaning Raw Data

We use a script file to be helpful of our data pre-processing task. This python script file is responsible for:

    i.   Remove all html tag name.

    ii.   Remove unnecessary spaces from the text.

    iii.   Remove all new line of each news and arrange it in a line.

    iv.   Assign a integer number for pre defining the category of each news.

This script result in a Tab Separated Value (tsv) formatted file and it looks like:

Figure 4.2.2: Tab separated Bangla text.

Actually, by this process, we can get all our categorical news in individuals file but the outputted file data are pre-processed and categorical.

4.4 Creating Input File

After data cleaning phase, we get six categorical tsv files as we are working on this research on these six categories. The nine categories are: Politics, Crime, Sports, Entertainment, Business, Life Style, Accident, National and International. Hence, after successfully preprocessing process, there have these six categorical news file in our hand. Then, to perform Natural Language Process on a Bangla news, we must join all these files into a file. For this, we use another python script named join.py. This file takes the folder name that contains all tsv files as an input and produces only a file where all news contained individually being merged.

4.5 Excluded Words Removal

We develop a python code for classify a news into a category. After joining all news into a file, our system is ready for building a model. For this, a little cleaning process is done before. We create a list that contains some Bangla words that actually no related with the category of the news. We called it as Excluded words and named it Excluded words list. Just checking that if excluded words are present in our input file or not. If exists, must be removed.

সেই	একটা	দুই	মনে	কাল						
ছিল	এ	এক	তাঁর	যে	আগে	আমার	বেশি	হবে	কিছু	কথা
নিয়ে	নিয়ে									
গত	সব	আমরা	এখন	খুব	আমাদের	শুরু	বছর	তো		
আরও										
হয়ে	করা									
হয়ে										
সেটা	মধ্যে	ভালো								
এটা	দিন	ওই								
অবশ্য	বড়	ভালো								
নতুন	শুধু	এবার	কাছে							
তাই										
খেলতে	কোনো	এবং	হতে	করেছেন						
সে	কি	তার	নিজের							
এর	পর্যন্ত									
তিন	আছে									
মতো	হবে	সালে	দুটি							

Figure 4.5.1: Bangla removed excluded text.

4.6 Feature Selection and Extraction

This phase is the main part of classifying approach and this is feature selection and extraction. It actually, decides, in which perspective classify will be done. We use word count as our feature selection and create it.

4.7 Building Model and Fit Dataset for Classifier

To build a model, we separate our dataset into two parts.

➢ Training Dataset
➢ Testing Dataset

We use 3:1 ratio for preparing our model. The three portion data set will be treated as training dataset and the rest one portion will be considered as testing dataset.
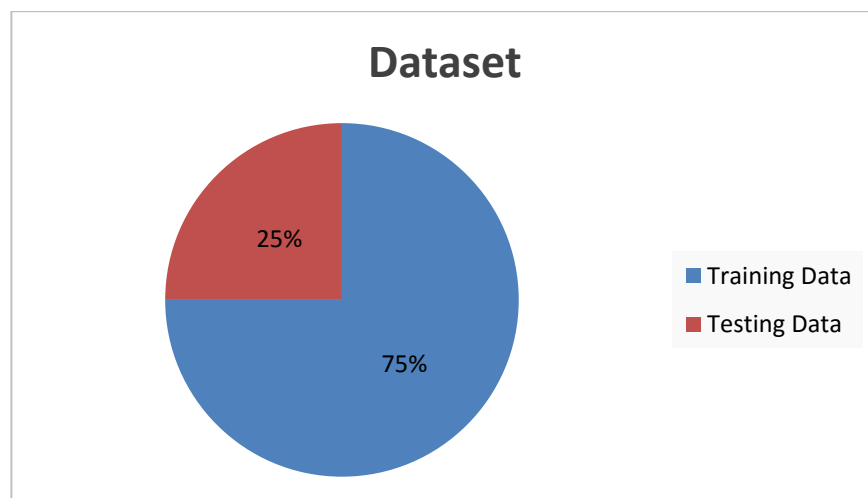

Figure 4.7.1: Dataset chart ratio.

In the concept of percentage, 75% data will be for training and 25% will be for testing. And this will make our expected model,

As, we are dealing with several classifier, we use it by importing sklearn package. This classifier can produce an integer that actually means the category of the expected news.

4.8 Experimental Result

After completing the classification of Bangla news, User Interface shown in figure 4.8.1 and figure: 4.8.2. It is an experimental input field where user can produce any kind of Bangla news text.

Figure 4.8.1: Graphical user interface.

Figure 4.8.2: Experimental output of Bangla news class "Sports".

Figure 4.8.3: Shows the experimental output of "Entertainment News".

4.9 Accuracy of Model

This is the Confusion Matrix of our model, confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in dataset.

For Naive Bayes Classifier

Table 4.9.1: Confusion Matrix for Naïve Bayes.

Output Input	Politics 0	Crime 1	Sports 2	Entertainment 3	Business 4	Life Style 5	Accident 6	National 7	International 8
Politics	141	28	4	0	5	0	0	12	12
Crime	10	206	0	0	2	0	0	17	3
Sports	13	1	383	10	1	0	0	12	4
Entertainment	3	01	10	77	0	0	0	12	3
Business	5	2	0	1	95	0	0	13	0
Life Style	2	0	0	0	2	5	0	4	0
Accident	0	28	0	0	0	0	1	1	0
National	28	43	3	12	21	1	0	298	8
International	5	9	6	2	6	0	0	12	50

Successfully Classified:

Table 4.9.2: Naive Bayes classified news type.

No.	News Type	Successfully Classify
1	Political News	141
2	Crime News	206
3	Sports News	383
4	Entertainment News	77
5	Business News	95
6	Life Style News	5
7	Accidental News	1
8	National News	298
9	International News	50
	Total	1256

Total News = 6530

Testing News (25%) = 1632.5

Accuracy of this model = (1256 / 1632.5) * 100

$$= 76.94\%$$

Table 4.9.3: Precision, recall, F1-Score for Naive Bayes.

Class Name	Precision	Recall	F1-Score
Politics	0.68	0.70	0.69
Crime	0.65	0.87	0.74
Sports	0.94	0.90	0.92
Entertainment	0.75	0.73	0.74
Business	0.72	0.83	0.77
Life Style	0.83	0.38	0.53
Accident	1.00	0.03	0.06
National	0.78	0.72	0.75
International	0.62	0.56	0.59
Average / Total	0.78	0.77	0.76

For K-Nearest Neighbors

Table 4.9.4: Confusion Matrix for K–Nearest Neighbors.

Output Input	Politics 0	Crime 1	Sports 2	Entertainment 3	Business 4	Life Style 5	Accident 6	National 7	International 8
Politics	97	26	41	2	0	0	0	22	10
Crime	8	126	85	0	0	0	1	17	1
Sports	1	4	404	8	0	0	0	6	1
Entertainment	0	2	31	53	0	0	0	20	0
Business	4	4	21	2	5	0	0	26	6
Life Style	1	0	1	0	1	5	0	5	0
Accident	1	11	13	0	0	0	5	0	0
National	16	30	123	12	9	1	5	211	7
International	1	8	40	3	2	0	0	12	24

Successfully Classified:

Table 4.9.5: K –Nearest Neighbors classified news type.

No.	News Type	Successfully Classify
1	Political News	97
2	Crime News	126
3	Sports News	404
4	Entertainment News	53
5	Business News	5
6	Life Style News	5
7	Accidental News	5
8	National News	211
9	International News	24
	Total	930

Total News = 6530

Testing News (25%) = 1632.5

Accuracy of this model = (930 / 1632.5) * 100

$$= 56.97\%$$

Table 4.9.6: Precision, Recall, F1-Score for K –Nearest Neighbors.

Class Name	Precision	Recall	F1-Score
Politics	0.75	0.48	0.59
Crime	0.60	0.53	0.56
Sports	0.53	0.95	0.68
Entertainment	0.66	0. 50	0.57
Business	0.77	0.46	0.57
Life Style	0.83	0.38	0.53
Accident	0.45	0.17	0.24
National	0.66	0.51	0.58
International	0.49	0.27	0.35
Average / Total	0.63	0.60	0.58
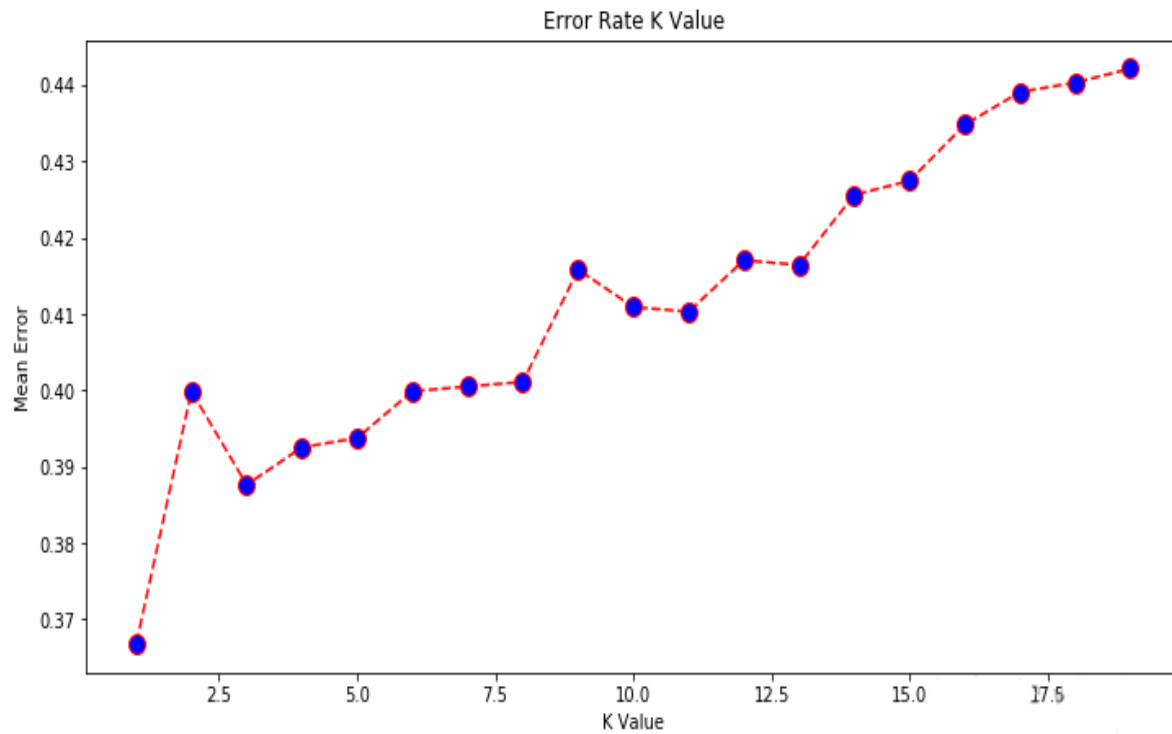
Figure 4.9.1: Error Rate of K Value.

For Decision Tree

Table 4.9.7: Confusion Matrix for Decision Tree

Output / Input	Politics 0	Crime 1	Sports 2	Entertainment 3	Business 4	Life Style 5	Accident 6	National 7	International 8
Politics	77	38	22	0	5	0	1	56	3
Crime	18	146	16	0	5	0	4	49	0
Sports	5	4	379	3	4	0	0	26	3
Entertainment	2	3	57	25	0	0	0	19	0
Business	3	8	14	0	39	0	0	49	3
Life Style	1	0	2	0	2	5	0	3	0
Accident	0	16	1	0	0	0	1	10	2
National	10	59	61	4	5	0	3	267	4
International	16	11	27	1	3	1	0	22	20

Successfully Classified:

Table 4.9.8: Decision Tree classified news type.

No.	News Type	Successfully Classify
1	Political News	77
2	Crime News	146
3	Sports News	379
4	Entertainment News	25
5	Business News	39
6	Life Style News	5
7	Accidental News	1
8	National News	267
9	International News	20
	Total	959

Total News = 6530

Testing News (25%) = 1632.5

Accuracy of this model = (959 / 1632.5) * 100

$$= 58.74\%$$

Table 4.9.9: Precision, Recall, F1-Score for Decision Tree.

Class Name	Precision	Recall	F1-Score
Politics	0.63	0.38	0.48
Crime	0.51	0.61	0.56
Sports	0.65	0.89	0.76
Entertainment	0.76	0. 24	0.36
Business	0.62	0.34	0.44
Life Style	0.83	0.38	0.53
Accident	0.11	0.03	0.05
National	0.53	0.64	0.58
International	0.57	0.22	0.32
Average / Total	0.59	0.59	0.56

For Random Forest

Table 4.9.10: Confusion Matrix for Random Forest.

Output Input	Politics 0	Crime 1	Sports 2	Entertainment 3	Business 4	Life Style 5	Accident 6	National 7	International 8
Politics	52	19	26	0	1	0	1	104	0
Crime	8	113	37	0	2	0	4	78	0
Sports	2	1	409	0	0	0	0	12	3
Entertainment	0	2	73	7	0	0	0	24	0
Business	9	3	12	0	20	0	0	72	0
Life Style	0	0	2	0	0	5	0	6	0
Accident	0	14	2	0	0	0	1	14	0
National	7	20	70	0	3	1	3	312	1
International	0	12	31	0	0	0	0	46	1

Successfully Classified:

Table 4.9.11: Random Forest classified news type.

No.	News Type	Successfully Classify
1	Political News	52
2	Crime News	113
3	Sports News	409
4	Entertainment News	7
5	Business News	20
6	Life Style News	5
7	Accidental News	1
8	National News	312
9	International News	1
	Total	920

Total News = 6530

Testing News (25%) = 1632.5

Accuracy of this model = (920 / 1632.5) * 100

$$= 56.35\%$$

Table 4.9.12: Precision, Recall, F1-Score for Random Forest.

Class Name	Precision	Recall	F1-Score
Politics	0.67	0.26	0.37
Crime	0.61	0.47	0.54
Sports	0.62	0.96	0.75
Entertainment	1.00	0. 07	0.12
Business	0.77	0.17	0.28
Life Style	0.83	0.38	0.53
Accident	0.00	0.00	0.00
National	0.47	0.75	0.58
International	0.50	0.01	0.02
Average / Total	0.50	0.56	0.50

For Support Vector Machine

Table 4.9.13: Confusion Matrix for Support Vector Machine.

Output Input	Politics 0	Crime 1	Sports 2	Entertainment 3	Business 4	Life Style 5	Accident 6	National 7	International 8
Politics	127	19	6	0	6	0	1	32	11
Crime	18	172	2	0	2	0	7	31	6
Sports	6	1	395	0	0	1	0	10	2
Entertainment	1	0	15	76	1	1	0	9	3
Business	3	2	1	1	92	0	0	15	2
Life Style	2	1	0	0	1	5	0	4	0
Accident	0	13	0	0	0	0	13	3	1
National	24	26	6	5	29	2	3	305	14
International	3	11	8	1	3	0	0	18	46

Successfully Classified:

Table 4.9.14: Support Vector Machine classified news type.

No.	News Type	Successfully Classify
1	Political News	127
2	Crime News	172
3	Sports News	395
4	Entertainment News	76
5	Business News	92
6	Life Style News	5
7	Accidental News	13
8	National News	305
9	International News	46
	Total	1231

Total News = 6530

Testing News (25%) = 1632.5

Accuracy of this model = (1231 / 1632.5) * 100

$$= 75.41\%$$

Table 4.9.15: Precision, Recall, F1-Score for Support Vector Machine.

Class Name	Precision	Recall	F1-Score
Politics	0.69	0.63	0.66
Crime	0.70	0.72	0.71
Sports	0.91	0.93	0.92
Entertainment	0.83	0. 72	0.77
Business	0.69	0.79	0.74
Life Style	0.56	0.38	0.45
Accident	0.54	0.43	0.48
National	0.71	0.74	0.73
International	0.54	0.51	0.53
Average / Total	0.75	0.75	0.75

Compare Algorithms

Table 4.9.16: Compare Precision of all classifier.

Algorithms	Accuracy
Naive Bayes	0.78
K–Nearest Neighbors	0.63
Decision Tree	0.59
Random Forest	0.60
Support Vector Machine	0.75

Table 4.9.17: Compare Recall of all classifier.

Algorithms	Accuracy
Naive Bayes	0.77
K–Nearest Neighbors	0.60
Decision Tree	0.59
Random Forest	0.56
Support Vector Machine	0.75

Table 4.9.18: Compare f1-score of all classifier.

Algorithms	Accuracy
Naive Bayes	0.76
K–Nearest Neighbors	0.58
Decision Tree	0.56
Random Forest	0.50
Support Vector Machine	0.75

Table 4.9.19: Compare algorithms accuracy.

Algorithms	Accuracy
Naive Bayes	76.94%
K–Nearest Neighbors	56.97%
Decision Tree	58.74%
Random Forest	56.35%
Support Vector Machine	75.41%

From the above comparison tables, we see that, in the case of Precision, Recall, f1-score and accuracy Naïve Bayes classifier is the best. The values of precision, recall and f1-score are respectively 0.78, 0.77, and 0.76 and the accuracy of this model is 76.94% that is highest value in comparison with all classifier.

4.10 Summary

After getting this accuracy, highest result come from Naïve Byes and Support Vector Machine that's why, we are satisfied, if we are try to increase accuracy level, must to prepare the dataset properly. The all categorical news should be equally numbered. At that, to increase the accuracy level, data cleaning has not alternative. The more data are preprocessed, the more accurate prediction will be shown by this classifier.

CHAPTER 5

SUMMAY, CONCLUSION, RECOMMENDATION AD IMPLECATION FOR FUTURE RESEARCH

5.1 Summary of the Study

It has no doubt that there are lots of research works on Natural Language Processing especially on English Language. When the outcome of such kind of works is taking a revolutionary change in our computing life, recently, such kind of research is being increased this time. We get some outstanding real life applications on the blessing of such kind of research works. But it is a matter of great regrets that there has no such of research work on Bangla Language. But it is the hope for us that many of researchers from the various countries have started to do research on this field. In our research work, we do some approaches of our Bangla News to classify its category.

5.2 Conclusion

Though, the accuracy level of the classifier algorithm that we used in our project is not so good but we have learnt lots of things from this research. We can now deal with the Bangla Text. We can now preprocess the row data. And can apply the classifier on our trained dataset. Hope, it will be very beneficial to the future researchers to do such kind of research on Bangla Text or Bangla news.

5.3 Recommendations

A few notable recommendations for this are as follows:

- To create the data set more efficiently, can produce a better output of this research work.

5.4 Implication for Further Study

- Adding more categories in this project, can make this more efficient.
- Using more classifiers on this dataset, can get a better understanding on which classifier can be the best for this work.

References

[1] Mansur, Mineral, "Analysis of n-gram based text categorization for Bangla in a newspaper corpus". Diss. BRAC University, 2006.

[2] Phani, Shanta, Shibamouli Lahiri, and Arindam Biswas, "A machine learning approach for authorship attribution for Bengali blogs." Asian Language Processing (IALP), 2016 International Conference on. IEEE, 2016.

[3] El-Barbary,O. G. El-Barbary, "Arabic news classification using field association words." SCIENCEDOMAIN Int 6.1 (1-9), 2016.

[4] Dutta, K., Kaushik, S. and Prakash, N, "Machine learning approach for the classification of demonstrative pronouns for Indirect Anaphora in Hindi News Items", The Prague Bulletin of Mathematical Linguistics, 95, pp.33-50 , Apr 2011 .

[5] Carreira, Ricardo, et al. "Evaluating adaptive user profiles for news classification." Proceedings of the 9th international conference on intelligent user interfaces. ACM, 2004.

[6] Haque, Rejwanul, et al. "English-Hindi transliteration using context-informed PB-SMT: the DCU system for NEWS 2009." Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration. Association for Computational Linguistics, 2009.

[7] Asy'arie, Arni Darliani, and Adi Wahyu Pribadi, "Automatic news articles classification in Indonesian language by using naive bayes classifier method." Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services. ACM, 2009.

[8] Buana, Putu Wira, and I. Ketut Gede Darma , "Combination of k-nearest neighbor and k-means based on term re-weighting for classify Indonesian news." International Journal of Computer Applications 50.11, 2012.

[9] Kanan, Tarek, and Edward A. Fox. "Automated arabic text classification with P-Stemmer, machine learning, and a tailored news article taxonomy." Journal of the Association for Information Science and Technology 67.11: 2667-2683, 2016.

[10] Kanan, Tarek, and Edward A. Fox, "Automated arabic text classification with P-Stemmer, machine learning, and a tailored news article taxonomy" Journal of the Association for Information Science and Technology 67.11: 2667-2683, 2009.

[11] Ee, Chee-Hong Chan Aixin Sun, and Peng Lim, "Automated online news classification with personalization." 4th international conference on asian digital libraries, 2001.

[12] Dilrukshi, Inoshika, Kasun De Zoysa, and Amitha Caldera, "Twitter news classification using SVM." Computer Science & Education (ICCSE), 2013 8th International Conference on. IEEE, 2013.

[13] Selamat, Ali, Hidekazu Yanagimoto, and Sigeru Omatu, "Web news classification using neural networks based on PCA." SICE 2002. Proceedings of the 41st SICE Annual Conference.,Vol. 4. IEEE, 2002.

[14] Kroha, Petr, and Ricardo Baeza-Yates, "A case study: News classification based on term frequency" Database and Expert Systems Applications, 2005. Proceedings. Sixteenth International Workshop on , IEEE, 2005.

[15] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 : 3-24, 2007.

[16] Billsus, Daniel, and Michael J. Pazzani, "A hybrid user model for news story classification." UM99 User Modeling. Springer, Vienna, 99-108 , 1999.

[17] Xu, Jun, Yu-Xin Ding, and Xiao-Long Wang, "Sentiment classification for Chinese news using machine learning methods." Journal of Chinese Information Processing 21.6 : 95-100, 2007.

[18] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 : 3-24 ,2007.

[19] Masand, Brij, Gordon Linoff, and David Waltz, "Classifying news stories using memory based reasoning." Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1992.

Appendix

Project Reflection

To complete the project we faced so many problem, first one was to determine the methodological approach for our project. It was not traditional work it was a research based project, more over there were not much work done before on this area. So we could not get that much help from anywhere. Another problem was that, collection of data, it was big challenge for us. There was no available source where we could get Bangla news text data, that's why we were develop a corpus for data collection. Also we started collect data manually. After a long time with hard work we could do that.

Plagiarism Report Screenshot:

Final _Reports.docx
1 hour ago

8%

Similarity

0%
Paraphrase

0%
Improper Citations

24
Matches

★ ★ ★
HIGH PLAGIARISM RISK