**BANGLA NEWSPAPER RECOMMENDATION SYSTEM BY MACHINE**

**LEARNING**

**by**

**MD. SHYKAT EMAM  
ID: 172-15-9669  
  
FAHIM RAHMAN TALUKDER  
ID: 172-15-10042**

This Report Presented in Partial Fulfillment of the Requirements for

The Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

**Ms. Zerin Nasrin Tumpa**

Lecturer

Department of Computer Science and Engineering

Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**18 MAY 2022**

**APPROVAL**

This Project titled “**BANGLA NEWSPAPER RECOMMENDATION SYSTEM BY MACHINE LEARNING**”, submitted by and 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. The presentation has been held on 18 may 2022.

**BOARD OF EXAMINERS**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

|  |  |
| --- | --- |
| **Dr. Touhid Bhuiyan**  **Professor and Head**  Department of Computer Science and Engineering  Faculty of Science & Information Technology  Daffodil International University | **Chairman** |

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

|  |  |
| --- | --- |
| **Name**  **Assistant Professor**  Department of Computer Science and Engineering  Faculty of Science & Information Technology  Daffodil International University | **Internal Examiner** |

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

|  |  |
| --- | --- |
| **Name**  **Lecturer**  Department of Computer Science and Engineering  Faculty of Science & Information Technology  Daffodil International University | **Internal Examiner** |

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

|  |  |
| --- | --- |
| **Name**  **Associate Professor**  Department of Computer Science and Engineering  United International University | **External Examiner** |

**DECLARATION**

We hereby declare that, this project has been done by us under the supervision of Name, Designation, Department of CSE Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

**Supervised by:**

**Ms. Zerin Nasrin Tumpa**

Lecturer

Department of CSE

Daffodil International University

**Submitted by:**

**MD. SHYKAT EMAM**

ID: 172-15-9669

Department of CSE

Daffodil International University

**FAHIM RAHMAN TALUKDER**

ID: 172-15-10042

Department of CSE

Daffodil International University

**ACKNOWLEDGEMENT**

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

We really grateful and wish our profound our indebtedness to Ms. Zerin Nasrin Tumpa, Lecturer, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Field name” to carry out this project. His endless patience, scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft 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. Touhid Bhuiyan**, 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**

The everyday newspaper is an especially imperative parcel of our life. It may take distinctive shapes, from the day by day everyday newspaper to pop-up notices when surfing the internet. The news is because it was given on an incidental premise. It is fundamental to communicate personalized news based on an individual's slants, personality, and excited condition. News can have a coincidental invaluable or appalling mental impact. With the improvement of advancement, time is getting to be a continuously limited resource. Machine learning and cognitive sciences can assist with anything from titles, organization, and planning to channel and personalization in news computing and preparation. Variety, coverage, uniqueness, and serendipity are all important aspects in a News Paper Recommendation System to provide a better user experience. In this work we tried to make machine learning based recommendation system. For this purpose, we utilized the tfidf technique to convert text to numbers for the algorithm learning process. For comparison and prediction, we use a total of five algorithms in this system: Logistic regression, Stochastic Gradient Descent, Random Forest, Support Vector Classifier, and Multinomial Nave Bayes. The maximum accuracy was around 90.29 percent, which was reached by employing a 30% data utilization rate by the Stochastic Gradient Descent method.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **CONTENTS** | **PAGE** |
| Board of examiners | i |
| Declaration | ii |
| Acknowledgements | iii |
| Abstract | iv |

|  |  |  |  |
| --- | --- | --- | --- |
| **CHAPTER** | | |  |
| **CHAPTER 1: INTRODUCTION** | | **1-4** | |
|  | * 1. Introduction   2. Motivation   3. Problem Definition   4. Research Question   5. Research Methodology   6. Research Objective   7. Summary | 01  02  02  03  03  03  04 | |

**CHAPTER 2: BACKGROUND 5-7**

2.1 Introduction 05

2.2 Related work 05

**CHAPTER 3: RESEARCH METHODOLOGY 9-12**

3.1 Introduction 09

3.2 Data Collection 10

3.3 Preprocessing 10

3.3.1 HTML Removing

3.3.2 Non Bangla Language Removing 11

3.4 Clean and Target Processing 12

3.4.1 Punctuation removing techniques 12

3.4.2 Target Manipulation 12

3.4.3 Down Sampling 14

3.4.4 Up Sampling 15

3.5 Dataset 16

3.6 Tokenization 16

3.7 Algorithm Implementation 17

3.7.1 Logistic Regression 17

3.7.2 SGD Classifier 18

3.7.3 Multinomial Naïve Bayes Classifier 18

3.7.4 Random Forest algorithm 18

3.7.5 Support Vector Classifier 19

3.8 Summary 19

|  |  |
| --- | --- |
| **CHAPTER 4: RESULT COMPARISON AND DISCUSSION** | **20-18** |
| 4.1 Experimental Setup | 20 |
| 4.2 Experimental Result and Analysis  4.3 Prediction Results Analysis for Existing Dataset  4.3.1 Random Forest  4.3.2 Multinomial Naïve Bayes  4.3.3 Support Vector Classifier  4.3.4 Logistic Regression  4.3.5 SGD classifier | 20  21  21  22  23  24  25 |

|  |  |
| --- | --- |
| **CHAPTER 5: Conclusion and Future Work** | **26-27** |
| 5.1 Introduction | 26 |
| 5.2 Conclusion | 26 |
| 5.3 Implication for Future Study | 26 |

|  |  |
| --- | --- |
| **REFERENCES** | **28** |
| **APPENDIX** | **29** |

|  |
| --- |
|  |
|  |
|  |
|  |
|  |

**LIST OF FIGURES**

|  |  |
| --- | --- |
| **FIGURES** | **PAGE NO.** |

|  |  |
| --- | --- |
| Figure 3.1: Methodology diagram | 9 |
| Figure 3.2: Preprocessing steps | 10 |
| Figure 3.3: All target variable | 12 |
| Figure 3.4: Selected Target | 13 |
| Figure 3.5: Imbalanced data rate | 14 |
| Figure 3.6: Down Sampling | 15 |
| Figure 3.7: up - sampling | 15 |
| Figure 3.8: Balance dataset | 16 |
| Figure 3.8 Logistic Regression | 17 |
| Figure 3.9: SGD classifier | 18 |
| Figure 3.5: Random Forest algorithm. | 19 |

**LIST OF TABLES**

|  |  |
| --- | --- |
| **TABLE NO.** | **PAGE NO.** |
|  |  |

|  |  |
| --- | --- |
| TABLE 3.1: HTML TAG REMOVING | 11 |
| TABLE 3.2: NON BANGLA REMOVING | 11 |
| TABLE 3.3: PUNCTUATION REMOVING | 12 |
| TABLE 3.4: TOKENIZATION | 16 |
| TABLE 3.5: PARAMETER USAGE | 17 |
| TABLE 4.1: ACCURACY COMPARISON | 20 |
| TABLE 4.2: RANDOM FOREST | 21 |
| TABLE 4.3: MULTINOMIAL NAÏVE BAYES | 22 |
| TABLE 4.4: SUPPORT VECTOR CLASSIFIER | 23 |
| TABLE 4.5: LOGISTIC REGRESSION | 24 |
| TABLE 4.6: SGD CLASSIFIER | 25 |
|  |  |
|  |  |
|  |  |

**CHAPTER 1**

**INTRODUCTION**

* 1. **Introduction**

Newspapers play a crucial role in our lives. It may take many forms, the daily newspaper to pop-up advertisements when surfing the web. The news is only sent on a regular basis. It is vital to provide customized news depending on a person's choices, mood, and emotional state. News can have an unintentionally positive or negative psychological influence. Time is becoming an increasingly scarce resource as technology advances. Machine learning and cognitive sciences can help with anything from title, organization, and scheduling to filtering and personalization in news computing and processing. Automatic newspaper headline suggestions may save you time while reading. It has previously been investigated, but all of the research has been conducted in English and is centered on the recommendation system of English newspapers, but no work has been completed in Bangla. In this research, we employed a large number of Bangla newspaper article to construct our data set because there has never been anything like it in Bangla before. A web crawler was used to acquire the Bangla data. Initially, we gathered around 15000 news articles. We chose 10000 news articles after preprocessing. A typical assessment criterion for determining the quality of a recommender system is accuracy, which was supplied by Prothomalo.com, dailynayadiganta.com, bd-pratidin.com, ittefaq.com.bd, jugantor.com, and samakal.com. Additional features such as variety, coverage, originality, and serendipity are crucial in providing a better user experience in a News Paper Recommendation System. Datasets, open recommendation platforms, and assessment methodologies all play a part in the development of news recommendation systems. Additional features such as variety, coverage, originality, and serendipity are crucial in providing a better user experience in a News Paper Recommendation System. Datasets, open recommendation platforms, and assessment methodologies all play a part in the development of news recommendation systems. For the algorithm learning process, we used the tfidf approach to transform text to numbers. This system employs five algorithms for comparison and prediction: logistic regression, stochastic gradient descent, random forest, support vector classifier, and multinomial Nave Bayes. The greatest accuracy was about 90.37 percent, which was achieved by using the Stochastic Gradient Descent technique with a data utilization rate of 30 percent.

* 1. **Motivation**

There are several newspaper recommendation systems accessible in English. However, there is no good recommendation system accessible for the Bangla language. This is the driving force behind our motivation. The recommendation system is becoming increasingly popular. Furthermore, recommendation systems allow users to make selections based on their intellectual preferences. Recommendation systems provide a one-of-a-kind and customized collection of information. The capacity of news websites to aid users in identifying interesting items to read is a significant challenge. In this post, we share our research into building a personalized news recommendation system in Bangla News. This was the main motivation of our work.

Reading time may be saved by using our solution by offering similar newspaper headline suggestions. For users who are signed in and have explicitly enabled browser history, the recommendation engine develops profiles of their news interests based on their past click behavior. We began by doing a large-scale study of anonymized Bangla news readers' click history to observe how their news interests change over time. Based on previous historical investigations, we developed a machine learning-based recommendation system.

**1.3 Problem Definition**

In today's world, machine learning is a very important issue in the field of natural language processing. Machine learning will assist in the advancement of the NLP business by recognizing several factors early on. We solved our problem step by step. The main problem was collecting data. we collected our data from different Bangla newspaper. The second problem is the collected data was not clean format there were lots of noise existed. So we need to clean them accurately as machine learning algorithm learn it very accurately. After preprocessing we analyzed the total dataset to acquire proper knowledge. At last we covert our dataset numerical format and then trained the algorithm.

**1.4 Research Questions**

The following are some of the issues addressed in this thesis:

* What is the data collection procedure?
* Is classification model being appropriate for this project?
* What are the data labeling procedure?
* How will this work help people?
* What is the most acceptable way to categorize news?
* Is it possible to real life integration of this project?

**1.5 Research Methodology**

Your research technique refers to how we intend to carry out your investigation. This chapter outlines how we intend to address issues such as data gathering methods, statistical analysis, participant observations, and so on. In this area of our study report, we collected the news Data Set, data cleaning and choosing characteristics, and feature selection techniques. After training the model, the result is determined using a classification approach (SVC, Logistic Regression, RF, SGD, MLP).

**1.6 Research Objectives**

The use of machine learning techniques in news classification has several advantages. There are several goals for employing machine learning techniques.

The following are some of the goals of machine learning techniques:

* Make a recommendation system based on Bangla news data.
* Try to save peoples time by providing appropriate suggestion.
* Try to find out which classification algorithm work better.
* Using an artificial intelligence-based approach, you may improve the accuracy of news suggestions.
  1. **Report Layout**

Chapter 1: will include the following topics: introduction, motivation, problem definition, research question, research methodology, and the anticipated end of our program. We also explain why we decided to perform this study in this section.

Chapter 2: The second chapter will look at the history of this study, as well as related studies and the current state of affairs from the perspective of Bangladesh. It includes a context analysis as well as a brief synopsis of the work.

Chapter 3: will explain the research technique This chapter delves deeply into the technique or workflow. This section will describe how the study was carried out.

Chapter 4: This chapter will go through the performance of the proposed model using an accuracy table and a classification report.

Chapter 5: This chapter is presented at the end of the report. The performance of the model is summarized in this section. This section also includes a comparison of accuracy. This section also discusses the web implementation and output of the model. The chapter concludes with a discussion of the work's shortcomings. It was also encoded with information about upcoming projects.

**1.8 Summary**

This chapter depicts the outline of our framework. This is our fundamental chapter. In this chapter, we give our overall framework outline as a presentation, along with a few related frameworks, inspirations, ambitions, and our commitments to this framework. This chapter looks at our overall framework strategy as well as how to solve our specific situation.

**CHAPTER 2**

**BACKGROUND**

**2.1 Introduction**

Nowadays, researchers of machine learning and artificial intelligence systems have come up with a slew of novel ideas for improving algorithms and their applications. Many projects are currently accessible on many websites throughout the world based on their planned system. The relevant work will be described in this chapter. We've supplied several recommendation system examples, as well as their performance. The existing system's accuracy is critical to how they use it, their forecast method and model, and where they put it in place.

**2.2 Related Works**

In recent years, the identification and recommendation of newspapers using past data has progressed using statistical and machine learning methodologies. For decades, CPD has been investigated in the fields of data mining, statistics, and computer science. In the following sections, we have covered most of the past research on recommendation engines in several disciplines utilizing machine learning, shallow learning, and deep learning.

In today's big data environment, news platforms highlight both the issues and opportunities for increasing the functioning of recommendation systems, according to Morteza et al. [6]. Using revolutionary big data storage and programming approaches, news recommendation systems can benefit from improved management and analysis of clickstream data, as well as a better understanding of users' preferences. To study user behavior, the bulk of existing news recommendation algorithms use users' clicks as implicit input.

Fuzzy logic, as defined by M.N.M Adnan et al., is a technique of calculation that use "degrees of truth" rather than the standard "true or false" (1 or 0) Boolean logic. User preferences are reflected by recommender systems, which then suggest things to read or browse based on those tastes. A variety of ways have been proposed for delivering suggestions, including content-based, collaborative, knowledge-based, and other approaches. Their method used fuzzy logic to find a cluster of articles that are related to one another and might be recommended to a reader. The application of fuzzy logic serves a straightforward function. The absolute terms of 0 and 1 are difficult to translate into news pieces that are relevant or recommendable. They can't merely point to an item called 'X' and say it has something to do with 'Y.' That's why they tried to develop a fuzzy algorithm based on a variety of features of a news item to determine whether or not an article is appropriate for user recommendation.

Traditional newspapers have been phased out in favor of digital media such as websites and purpose-built mobile applications, according to C. Feng et al. [2]. News recommendation systems have been shown to be capable of automatically assessing long articles and recommending similar items to customers based on pre-defined criteria. The writing is analyzed in its entirety from 2001 to 2019, and 81 connected considerations are nominated, essentially grouped into six categories, and discussed in this work. According to the findings, 60 percent of news proposal frameworks adopted a hybrid approach, 66 percent considers tiny conversations almost datasets, and many concerns from a lengthy list of issues in the news arena are addressed. This is the first piece in the area to present a complete large-scale picture of news suggestion and to analyze various measures found inside the ponders. The last section contains long-term research opportunities that will lead to progress in the proposal of news pieces inside the news space.

A social network model of a trust-based recommendation system is presented by Walter et al. [7]. Agents access information via their social network and filter it through their trust links, according to the model's concept. They compared the system's performance to that of a frequency-based recommendation system to discover how agent trust dynamics influence the system's success. They also discovered that network density, preference heterogeneity among agents, and knowledge sparseness are all significant factors in the system's success. The system self-organizes into a near-optimal state; the system's global performance is an emergent property, resulting from local agent interactions rather than purposeful coordination.

H. Tan and colleagues. [3] When learners lack appropriate personal experience with the possibilities, e-learning recommendation systems can help them make decisions, which is crucial in this age of information overload. The primary recommendation method in our study, which is linked with online education, is the user-based collaborative filtering technique. Data collection, data ETL, model construction, strategy setup, and service supply are the five phases of a web-based e-learning recommendation system's workflow. A basis for future expansion is also given in the form of an architecture. This architecture consists of seven modules, four of which are essential: recommendation models database, recommendation system database, recommendation administration, and data/model management.

Goynai et al. [4] in this age of information, an e-learning recommendation system can help learners make judgments when they don't have enough personal experience with the possibilities. The primary recommendation method in our study, which is linked with online education, is the user-based collaborative filtering technique. Data collection, data ETL, model construction, strategy setup, and service supply are the five phases of a web-based e-learning recommendation system's workflow. A basis for future expansion is also given in the form of an architecture. Suggestion models database, recommendation system database, recommendation management, and data/model management are the four major modules in their architecture.

The Netflix Prize, according to Gábor et al.[8], is a collaborative filtering problem. The emergence of online companies that integrated recommendation systems propelled the area of machine learning to prominence in the late 1990s. The goal of such a system is to forecast what a user would appreciate based on prior ratings and those of other users. Because the Netflix Prize dataset is far larger than previous benchmark datasets, the methodologies must be scalable. The major components of their blending-based technique, dubbed the Gravity Recommendation System, are explained in this study (GRS). As of November 2007, it has an RMSE of 0.8743 in the Netflix Prize competition.

In this age of information overload, Alejandro et al. [5] employ an E-learning recommendation system to aid learners in making judgments when they do not have enough personal experience with the possibilities. The primary recommendation method in our study, which is linked with online education, is the user-based collaborative filtering technique. Data collection, data ETL, model construction, strategy setup, and service supply are the five phases of a web-based e-learning recommendation system's workflow. A basis for future expansion is also given in the form of an architecture. This architecture has seven modules, four of which are essential: suggestion models database, recommendation system database, recommendation management, and data/model management.

Reza et al. [9] illustrate that an increasing number of people are reading news online, where they have access to millions of articles from a variety of sources. News recommender systems (NRS) were created to help users locate the proper and relevant material and to alleviate the problem of information overload by suggesting news articles that would be of interest to news readers. They identified alternative solutions from the state-of-the-art in this study, which emphasized the primary issues encountered by the NRS. Their debate is split into two halves. They give a summary of the recommendation solutions, datasets, assessment criteria other than accuracy, and recommendation platforms utilized in the NRS in the first half.

**CHAPTER 3**

**RESEARCH METHODOLOGY**

**3.1 Introduction**

I'll take you through the stages of our unique Bengali news classification technique in this section. The selected model includes data collection, data cleaning, and data pre-processing, as well as accompanying equations, diagrams, figures, tables, and explanations. The following is a detailed description of the system design we suggest. Discuss the architecture's general details. We walk through and demonstrate each aspect of our proposed system design. Discuss the recommended system's algorithm and operation:

Data Collection

Preprocessing

Dataset

Tokenization

Algorithm Implementation

Evaluation

RF

LR

SGD

MNB

SVC

HTML Removing

Non Bangla Remoing

Clean and Target Process

Figure 3.1: Methodology diagram

**3.2 Data Collection**

In addition to data collection, data was acquired from a number of well-known Bengali newspapers, and a massive dataset was built by integrating all of the newspapers and classifying them with their respective categories. I also acquired data on the data source and the title of the content. I had a large number of datasets to deal with, so gathering data from a variety of sources was tough. We obtained the information we required using the kaggle website. This is a real dataset from a Bangladeshi newspaper that was crawled. There is a significant amount of newspaper content preserved there. Some examples are dailynayadiganta.com, prothomalo.com, bd-pratidin.com, jugantor.com, manobkantha.com.bd, ittefaq.com.bd, dailyjanakantha.com, jaijaidinbd.com, and dhakatribune.com. We started with a total of 20000 data points.

**3.3 Preprocessing**

In our preprocessing stage, there are three stages that depart. Html Tag Removal, Non-Bangla Tag Removal, Data Cleaning, and Target Manipulation are all examples of HTML Tag Removal.

dailynayadiganta.com, prothomalo.com, bd-pratidin.com, jugantor.com, manobkantha.com.bd, ittefaq.com.bd, dailyjanakantha.com, jaijaidinbd.com, and dhakatribune.com are some of the websites that provide information. Initially, we gathered 20000 data points.

HTML Removing

Non Bangla Removing

Clean and Target Process

Figure 3.1: Preprocessing steps

There are three main steps of preprocessing steps. Html Removing, Non Bangla Removing, Clean and Target Process. Each of the process description is given bellow:

**3.3.1 HTML Removing:**

Table 3.1: Html Tag Removing

|  |  |
| --- | --- |
| Raw Text | Only Bangla Text |
| '<p><span>Guided writing \uf06e</span></p>', '<p>চাঁদপুরের হাজীগঞ্জ পৌরসভাধীন বলাখাল এলাকায় আবদুর রহি | চাঁদপুরের হাজীগঞ্জ পৌরসভাধীন বলাখাল এলাকায় আবদুর রহি |
| '<p><strong>এটিএন বাংলা</strong><br />সকাল ১০-৪০ ব্যথার দান (শাবানা, আলমগীর, দিলারা)।<br /><strong>এনটিভি</strong><br /> | এটিএন বাংলাসকাল ১০-৪০ ব্যথার দান (শাবানা, আলমগীর, দিলারা)। |

For html removing we used lxml library. By this library we removed <p>, <h1>, <div> tag. As these tag is not necessary for our model we removed this. Table 3.1 represents the html tag removing.

**3.3.2 Non Bangla Language Removing:**

Table 3.2: Non Bangla Removing

|  |  |
| --- | --- |
| Raw Text | Only Bangla Text |
| প্রিয় শিক্ষার্থীরা, আজ ইংরেজি ২য় পত্রের Part-A-‰i Transformation of sentences নিয়ে আলোচনা করব | প্রিয় শিক্ষার্থীরা, আজ ইংরেজি ২য় পত্রের নিয়ে আলোচনা করব |
| &gtআজ ১৩ আগস্ট। চার বছর আগে এই দিনে নির্মাতা তারেক মাসুদ | আজ ১৩ আগস্ট। চার বছর আগে এই দিনে নির্মাতা তারেক মাসুদ |
| A prodigal man is not as happy as a frugal man. সহকারী অধ্যাপক ঢাকা স্টেট কলেজ, ঢাকা | সহকারী অধ্যাপক ঢাকা স্টেট কলেজ, ঢাকা |
| নিয়মিত চীনাবাদাম খেলে ক্যানসার ও হৃদ্\u200cরোগে অকালমৃত্যুর ঝুঁকি কমে | নিয়মিত চীনাবাদাম খেলে ক্যানসার ও হৃদ্ রোগে অকালমৃত্যুর ঝুঁকি কমে |

Table 3.2 represents the non-Bangla removing process. In data preprocessing process we have seen that there are some non-Bangla language like English and other language available some article. As we work only Bangla language so I need to remove all non bangla language. And we successfully removed all non-Bangla languages by regular expression.

**3.4 Clean and Target Processing**

The data clean and target processing is a very complex process in our work, we will explain each step one by one below:

**3.4.1 Punctuation removing techniques**

Table 3.3: Punctuation Removing

|  |  |
| --- | --- |
| Raw Text | Clean Text |
| সিরাজগঞ্জ-৩ (রায়গঞ্জ-তাড়াশ) আসনের উপনির্বাচনে মনোনয়নপত্র জমা দিতে গিয়ে সম্ভাব্য এক প্রার্থী আওয়ামী লীগের নেতা-কর্মীদের হাতে লাঞ্ছিত হয়েছেন বলে অভিযোগ পাওয়া গেছে। এ সময় ছবি তুলতে গিয়ে স্থানীয় এক সাংবাদিকও লাঞ্ছিত হন। গতকাল বৃহস্পতিবার সকালে উপজেলা পরিষদে এ ঘটনা ঘটে। | সিরাজগঞ্জ ৩ রায়গঞ্জ তাড়াশ আসনের উপনির্বাচনে মনোনয়নপত্র জমা দিতে গিয়ে সম্ভাব্য এক প্রার্থী আওয়ামী লীগের নেতা-কর্মীদের হাতে লাঞ্ছিত হয়েছেন বলে অভিযোগ পাওয়া গেছে। এ সময় ছবি তুলতে গিয়ে স্থানীয় এক সাংবাদিকও লাঞ্ছিত হন গতকাল বৃহস্পতিবার সকালে উপজেলা পরিষদে এ ঘটনা ঘটে |

Table 3.1 shows the process of removing punctuation marks. We decided to remove the penalty since understanding the spirit of a statement is not required.

**3.4.2 Target Manipulation**

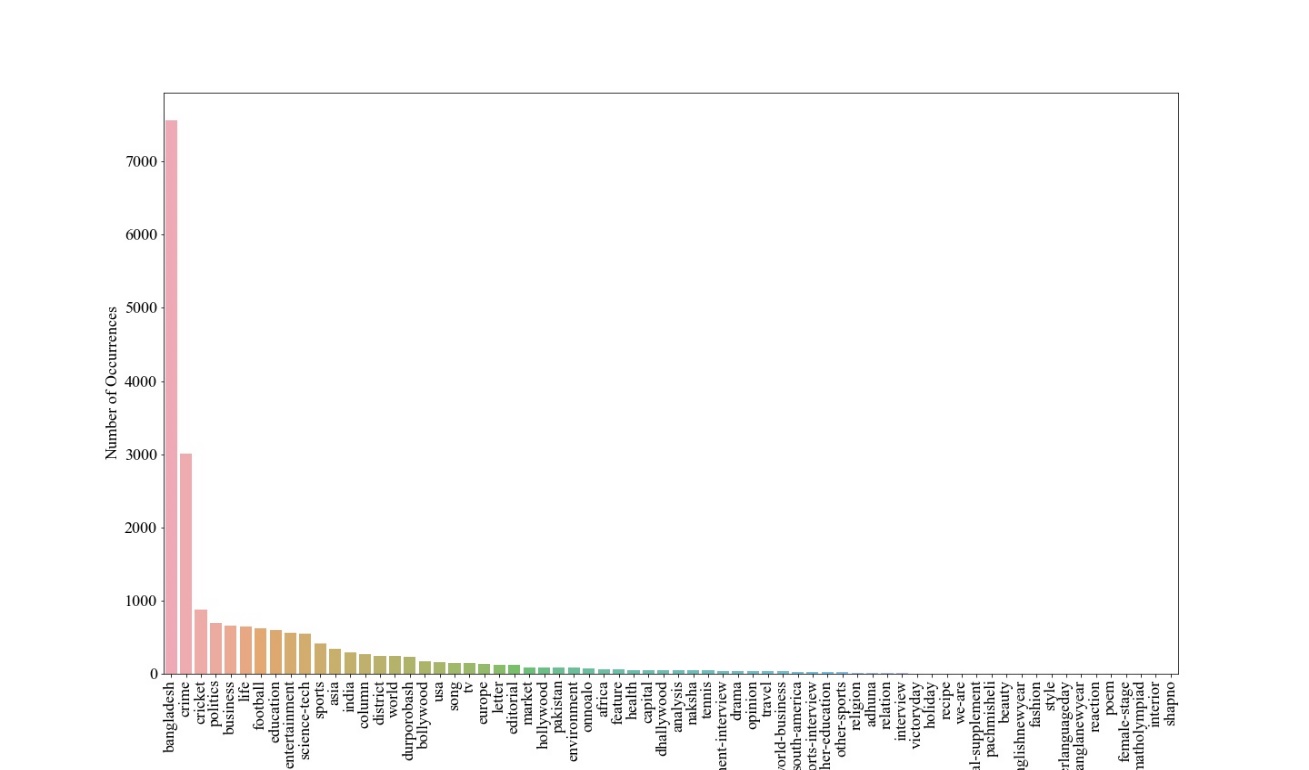
In this section, we looked at a lot of data. Each aspect of the analysis is listed below.

Figure 3.3: All target variable

Figure 3.3 depicts the initial target variable of our study. There were around 68 unique targets among the 20,000 data points we collected. Unfortunately, the vast majority of them are outliers. This implies that the great majority of the target population lacks the necessary data to train properly. As a result, we've decided to rule out any and all outliers. As a consequence, we've opted to keep a target that has at least 500 data rows.

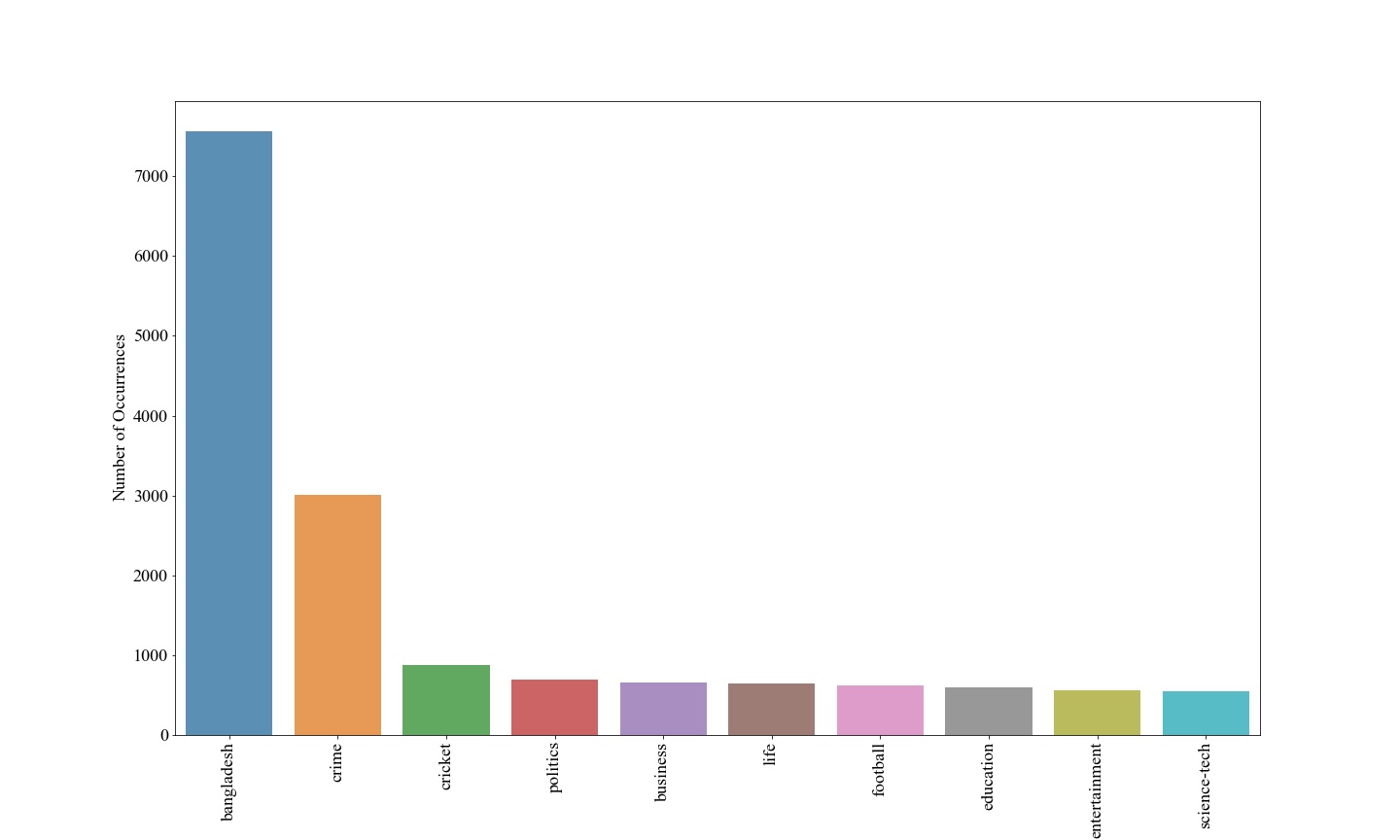


Figure 3.4: Selected Target

Figure 3.4 depicts the selected target. The 10 objectives include Asia, Bangladesh, business cricket, crime, education, entertainment, football, life, politics, and science and technology. Figure 3.3 shows all of the targets in descending order. Bangladesh has the greatest data, with around 8000 rows. The least amount of data the target includes is roughly 800. Our dataset is completely uneven, as we can see from the description above. This dataset should not be trained directly.

Figure 3.5 represents the Imbalanced percentage rate of our selected target.

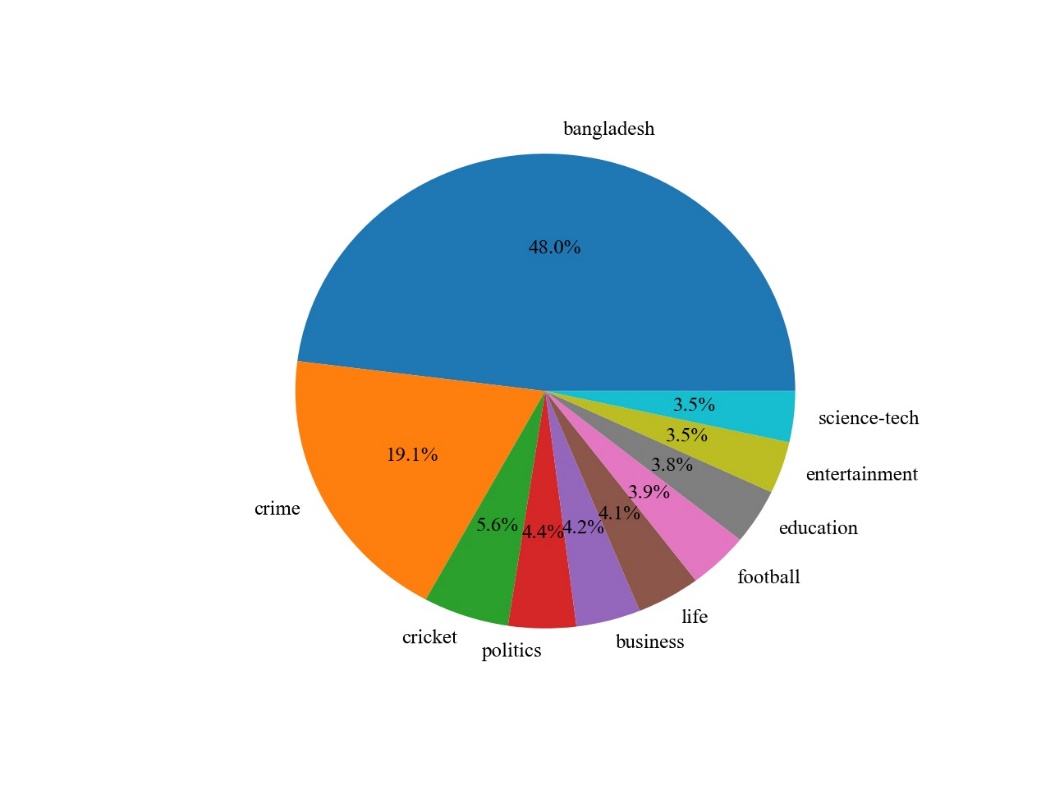
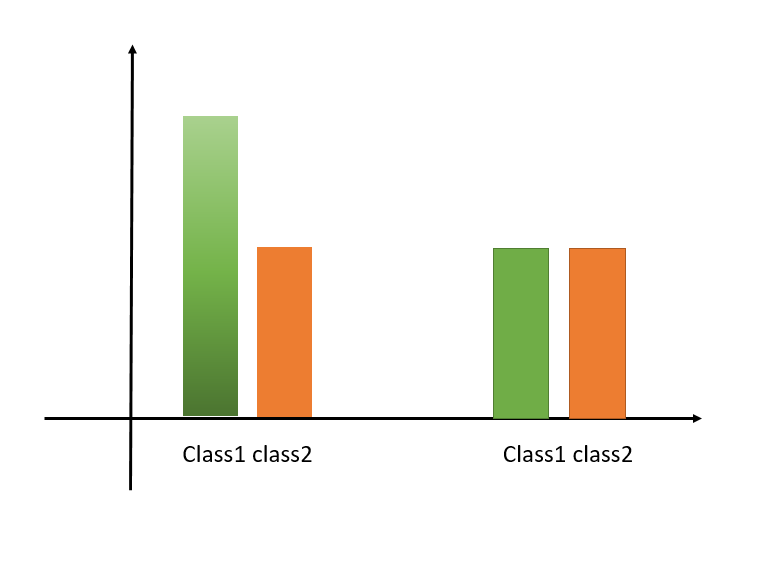
Figure 3.5: Imbalanced data rate

Figure 3.5 represents the imbalanced data rate percentage. The highest percentage is contained by Bangladesh with percentage rate 48.0%. and the lowest percentage is 3.5% acquired by science-tech. from this graph we can easily say that the dataset is very imbalanced and unqualified for training machine learning algorithm. So we need to balance this dataset. There are two different popular ways to balance dataset. one is up sampling another is down sampling. In our work we used both of techniques.

**3.4.3 Down Sampling**

Down examining could be a procedure for bringing down the sum of preparing tests from the larger part course. Since it contributes to the by and large adjust of the objective category. In down-sampling, we haphazardly evacuate the perceptions from the lion's share course. In this way after up-sampling or down-sampling, the dataset gets to be adjusted with same number of perceptions in each lesson. When we eradicate procured information, we ordinarily lose a parcel of imperative data. Figure 3.6 represents the down sampling process.

Figure 3.6: Down Sampling

**3.4.4 Up Sampling**

The method of infusing zero-valued tests between unique tests in arrange to improve the testing rate is known as up testing. This sort of up examining presents undesirable ghastly pictures into the first flag, which are centered on products of the testing rate. Figure 3.6 speak to up inspecting prepare.

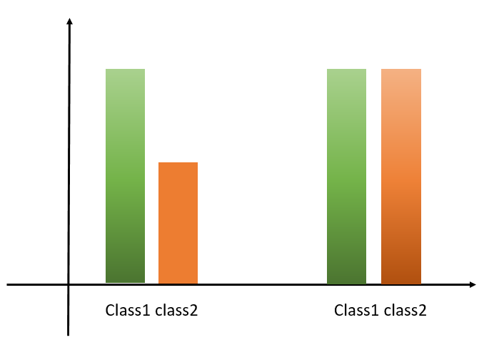
Tragically, both of these prepare contains stars and cons. For up examining cons is information will be copy. And for down inspecting information never copy but data will be misfortune. So for appropriate adjusting we actualize both of the method at a time. Which means a few information we connected up testing and for a few information we have actualize down sampling prepare. As a result, we got the correct adjust dataset. And each of the target contains break even with number of information.

Figure 3.7: up - sampling

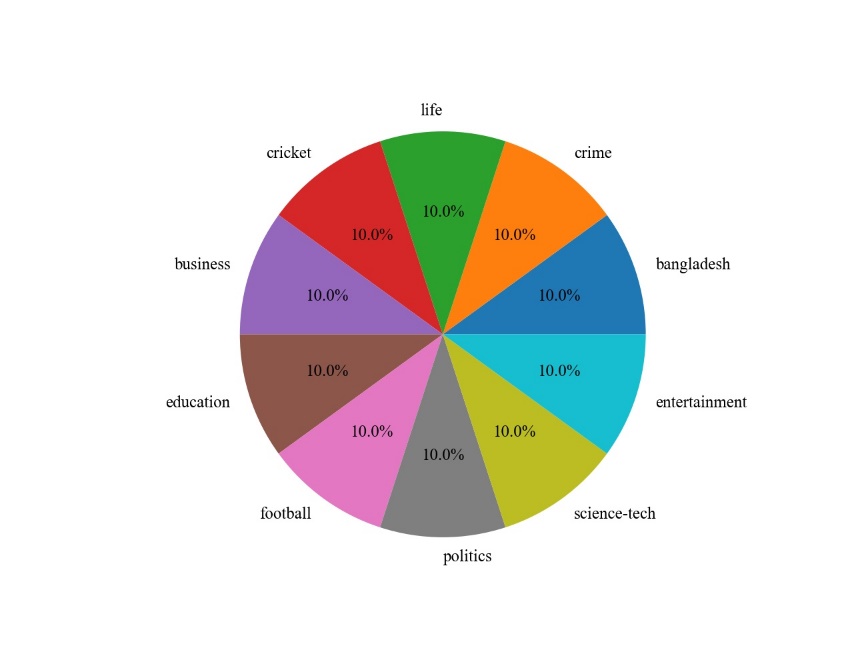
**3.5 Dataset**

Figure 3.8: Balance dataset

Figure 3.8 represents the balance dataset. this dataset is our final dataset by applied up sampling and down sampling we got this dataset. In this dataset each of the class contains 10.0% data of 10000 data. for further work we used this dataset.

**3.6 Tokenization**

Tokenization is the method of changing over touchy data into no sensitive "tokens" that will be utilized in a database or inside model without bringing it into scope. [10] In spite of the reality that the tokens are disconnected values, they keep up a few angles of the initial information, such as length or arrange, permitting them to be utilized in trade exercises without intrusion. The initial touchy data is at that point safely kept exterior of the company's possess model.

Table 3.4: Tokenization

|  |  |
| --- | --- |
| Input data | Tokenized data |
| প্রিয় শিক্ষার্থীরা, আজ ইংরেজি ২য় পত্রের নিয়ে আলোচনা করব | ‘প্রিয়’, ‘শিক্ষার্থীরা,’, ‘আজ’, ‘ইংরেজি’ ,‘২য়’, ‘পত্রের’ ‘নিয়ে’ ,‘আলোচনা’, ‘করব’ |
| আজ ১৩ আগস্ট। চার বছর আগে এই দিনে নির্মাতা তারেক মাসুদ | ‘আজ’ ,‘১৩’, ‘আগস্ট’ ,‘চার’, ‘বছর’, ‘আগে’, ‘এই’, ‘দিনে’, ‘নির্মাতা’ ,‘তারেক’ ,‘মাসুদ’ |
| সহকারী অধ্যাপক ঢাকা স্টেট কলেজ, ঢাকা | ‘সহকারী’, ‘অধ্যাপক’ ,‘ঢাকা’ ,‘স্টেট’, ‘কলেজ’ ,‘ঢাকা’ |

Tokenized data is shown in Table 3.2. We construct a token for each phrase as we work at the word level.

**3.7 Algorithm Implementation**

In computer science, implementation refers to the programming and deployment of a technical specification or technique into a program, software component, or other computer system. A given specification or standard may have many implementations. There were two approaches to implementing the procedure. Machine learning methods are used in shallow learning strategies. Stochastic gradient descent, random forest, support vector classifier, and multinomial Nave Bayes are examples of logistic regression algorithms.

Table 3.5: Parameter usage

|  |  |
| --- | --- |
| Algorithms | Details |
| Logistic Regression | fit\_intercept = True , tol = 1e-4, penalty = l1 |
| Multinomial Naïve Bayes Classifier | Alpha = 1.0, fit\_prior = True, class\_prior = None |
| Stochastic gradient descent | fit\_intercept = True , penalty = l1, shuffle = True |
| SVC | Kernel = linear, degree = 3 |
| Random Forest | n\_estimators=100, min\_samples\_leaf = 1 |

**3.7.1 Logistic Regression**

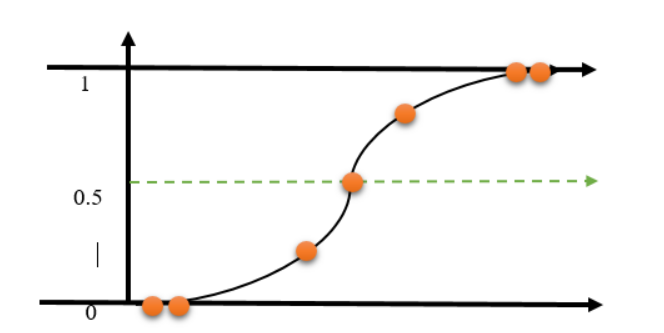
 Logisticregression is a machine learning technique for classification issues that is based on the probability notion. Logistic regression models are similar to linear regression models in that they employ more sophisticated cost functions, known as "sigmoid functions" or "logistic functions," rather than linear functions. A logistic regression model is depicted in Figure 3.10.

Figure 3.8 Logistic Regression

**3.7.2 SGD Classifier**

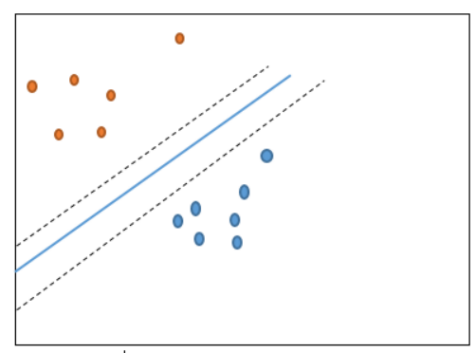
To assess the performance of your dataset, use the sgd classifier. SGD is a straightforward and effective method for adapting linear classifiers and regressionrs to convex loss functions, such as (linear) support vector machines and logistic regression. Stochastic Gradient Descent (SGD) is a straightforward yet effective optimization method for identifying the values of function parameters / coefficients that minimize the cost function. In other words, it is utilized for discriminative learning of linear classifiers employing convex loss functions like SVM and logistic regression. Because the coefficients are updated each training instance rather than at the end of the sample, they may be utilized on big datasets with success. The sgd classifier is seen in Figure 3.9.

Figure 3.9: SGD classifier

**3.7.3 Multinomial Naïve Bayes Classifier**

This is a classification approach based on Bayes' theorem and the predictor-independence assumption. Simply expressed, the naive Bayes classifier believes that the existence of one function in a class is unrelated to the existence of another. Although Naive Bayes is a basic algorithm, it can outperform a more complicated classification system. The naive Bayes classification algorithm is represented by Equation 1 [11].

(1)

**3.7.4 Random Forest algorithm**

Random forest is a supervised machine learning technique for solving classification problems. Like decision trees, it supports both categorical and continuous input and output variables. In the CART model, a single tree grows in a random forest and multiple trees grow in a random forest. To classify new objects based on quality, each tree assigns a classification. This is called a tree "vote" for that class. In regression, Forest selects the category with the most votes and averages the results of many trees. [12].

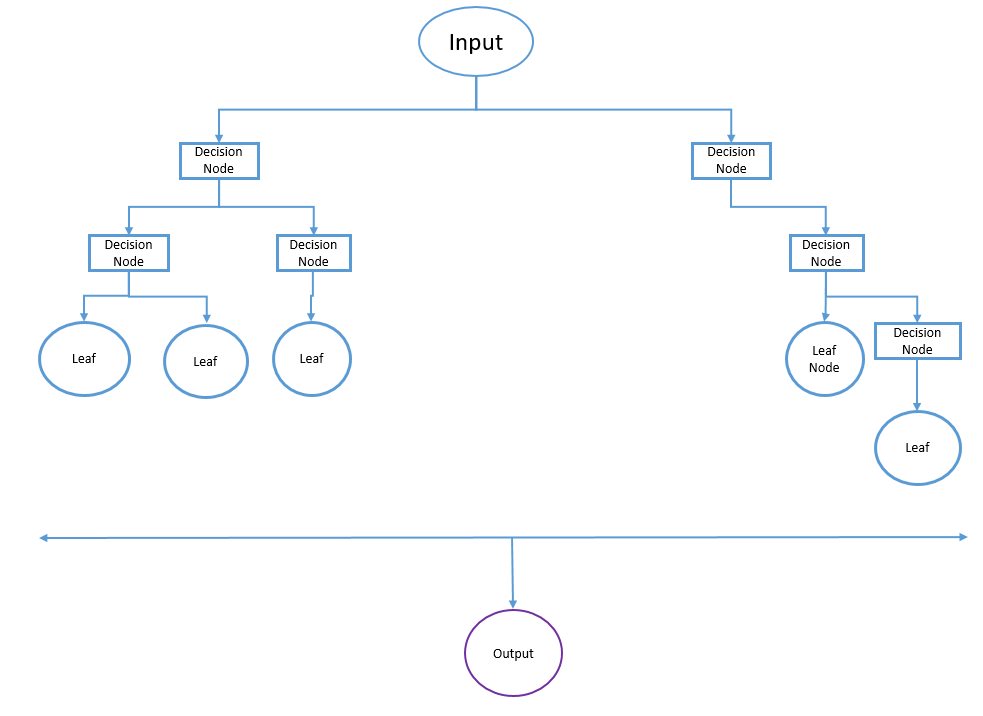


Figure 3.5: Random Forest algorithm.

**3.7.5 Support Vector Classifier**

Back Vector Machines (SVMs) can be a machine learning innovation related to classification and fallback issues. However, it is most often used for classification issues. In the SVM calculation, each piece of information is represented as a point in n-dimensional space, and the weight of each peak is the weight of a particular placement. At this point, select the hyperplane that best recognizes the two classes to classify the information. The SVM strategy ignores exceptions and provides inclusion that allows you to distinguish the hyperplane with the highest edges. As a result, SVM classification can be claimed to be safe against exceptions. Within the SVM classifier, it is clear to set the hyperplane directly between these two classes. Anyway, another important point is the possibility that this ability should be physically consolidated in order to have a hyperplane. The SVM approach, on the other hand, adheres to a technique called bit trapping. SVM bits may be a job to come with

**3.8 Summary**

In this chapter we have outlined our techniques and the calculations we used in our work. We also outline the scientific instincts of each calculation and address them all graphically. This chapter describes the general workflow and flow charts for this technique. we completed them one by one.

**CHAPTER 4**

**Experimental Results and Discussion**

**4.1 Experimental Setup**

In the breakdown, we first present the current data set's prediction results, followed by the results of our own forecasts regarding the data set. We tried a lot of algorithms and determined that five of them were the best. Precision, accuracy, reminder, and f1 were also chosen as data computation parameters. We utilized dataset slice styles to find the accuracy. The first is a dataset containing a prevent phrase, while the second is a dataset without a prevent phrase. This strategy was used to determine which dataset is better for our model. We employed a different strategy for each dataset, which is the check facts usage percent rate. We utilized a check facts utilization rate of 25 to 40 percent. When check is 25%, then educate can be 75% of our general facts. In the same way, if check facts is 40%, then educate can be 60%. This approach was utilized to determine which percentage is higher for our algorithm.

**4.2 Experimental Result and Analysis**

For prediction, logistic regression, stochastic gradient descent, random forest, support vector classifier, and naive Bayes were employed. On the basis of accuracy and accuracy score, we compared these algorithms. The Random Forest and Decision Tree algorithms reach 90.29 percent accuracy with a 25 percent usage rate, according to the accuracy table.

Table 4.1: Accuracy Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test data usage rate | | 25% | 30% | 35% | 40% |
| Algorithms  Accuracy | *RF* | 85.74 | 85.00 | 84.32 | 83.19 |
| *MNB* | 86.75 | 86.78 | 86.02 | 85.15 |
| *SVC* | 88.52 | 88.05 | 87.74 | 87.31 |
| *Logistic* | 87.96 | 87.99 | 87.05 | 86.66 |
| *SGD* | 90.29 | 89.56 | 88.95 | 88.40 |

the results of the tests to decide which was the most effective The yellow box shows the proportion of trials for each algorithm that obtained the maximum accuracy. The most accurate classifier is the SGD classifier. As a result, if only 25% of the test data is used, the accuracy rate will be 90.29 percent. The results of a 25% random forest, a Nave Bayes polynomial, a support vector classifier, and the Logistic Regression approach are 85.74, 86.75, 88.52, and 87.96, respectively.

**4.3 Prediction Results Analysis for Existing Dataset**

In this part, we can pass over the numerous rating matrix analyses of every method. As maximum of the set of rules generates excellent accuracy the usage of 25% take a look at data. As a result, we performed a rating matrix evaluation using 25% of the data. The evaluation is furnished below:

**4.3.1 Random Forest**

From Table 5.1, you can see the score matrix table for the Random Forest algorithm. For column data Random Forest achieved the highest score. The score is 0.97 for memory 0.98 and 91 for targeting, soccer and ScienceTech ScienceTech..

Table 4.2: Random Forest

**Table 5.1: Random Forest**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Target** | **Precision** | **Recall** | **F1-Score** | **Support** |
| bangladesh | 0.66 | 0.51 | 0.58 | 237 |
| business | 0.85 | 0.83 | 0.84 | 270 |
| cricket | 0.9 | 0.9 | 0.9 | 242 |
| crime | 0.77 | 0.91 | 0.84 | 249 |
| education | 0.98 | 0.93 | 0.95 | 263 |
| entertainment | 0.94 | 0.94 | 0.94 | 243 |
| football | 0.82 | 0.97 | 0.89 | 247 |
| life | 0.95 | 0.75 | 0.84 | 250 |
| politics | 0.83 | 0.91 | 0.87 | 251 |
| science-tech | 0.89 | 0.93 | 0.91 | 248 |
|  |  |  |  |  |
| Accuracy |  |  | 0.86 | 2500 |
| Macro AVG | 0.86 | 0.86 | 0.86 | 2500 |
| Weighted AVG | 0.86 | 0.86 | 0.86 | 2500 |

**4.3.2: Multinomial Naïve Bayes**

Table 5.2 represents the multinomial naive Bayes algorithm. This algorithm works with so many functions that it is the most used algorithm in NLP. However, unlike the Random Forest algorithm, it only gets the highest score on soccer and educational goals. The percentages are 98 and 97.

Table 4.3: Multinomial naïve bayes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Target** | **Precision** | **Recall** | **F1-Score** | **Support** |
| bangladesh | 0.77 | 0.47 | 0.59 | 237 |
| business | 0.86 | 0.92 | 0.89 | 270 |
| cricket | 0.96 | 0.96 | 0.96 | 242 |
| crime | 0.82 | 0.85 | 0.84 | 249 |
| education | 0.97 | 0.92 | 0.95 | 263 |
| entertainment | 0.96 | 0.93 | 0.94 | 243 |
| football | 0.98 | 0.97 | 0.97 | 247 |
| life | 0.89 | 0.84 | 0.87 | 250 |
| politics | 0.71 | 0.96 | 0.82 | 251 |
| science-tech | 0.89 | 0.95 | 0.92 | 248 |
|  |  |  |  |  |
| Accuracy |  |  | 0.88 | 2500 |
| Macro AVG | 0.88 | 0.88 | 0.87 | 2500 |
| Weighted AVG | 0.88 | 0.88 | 0.88 | 2500 |

**4.3.3: Support Vector Classifier**

The most reached peak of 0.98 is generated by the support vector classifier. In terms of accuracy, we have achieved the highest value of 0.98 in soccer. For recall and f1 result support vector classifiers, the maximum soccer goal variables were 0.97 and 0.98. Macro AVG and weighted AVG are very suitable for accurate search and f1 score.

Table 4.4: Support Vector Classifier

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Target** | **Precision** | **Recall** | **F1-Score** | **Support** |
| bangladesh | 0.62 | 0.68 | 0.65 | 237 |
| business | 0.89 | 0.89 | 0.89 | 270 |
| cricket | 0.96 | 0.97 | 0.96 | 242 |
| crime | 0.84 | 0.85 | 0.85 | 249 |
| education | 0.97 | 0.92 | 0.95 | 263 |
| entertainment | 0.95 | 0.95 | 0.95 | 243 |
| football | 0.98 | 0.97 | 0.98 | 247 |
| life | 0.88 | 0.85 | 0.86 | 250 |
| politics | 0.9 | 0.88 | 0.89 | 251 |
| science-tech | 0.92 | 0.93 | 0.93 | 248 |
|  |  |  |  |  |
| Accuracy |  |  | 0.89 | 2500 |
| Macro AVG | 0.89 | 0.89 | 0.89 | 2500 |
| Weighted AVG | 0.89 | 0.89 | 0.89 | 2500 |

**4.3.4: Logistic Regression**

Under five different machine learning algorithms, the logistic algorithm produces 88% accuracy. There is no maximum value of 97 generated by the logistic algorithm. The maximum accuracy of entertainment and soccer is 0.97. Support 288 or 305 with support. Macro and weighted averages are also much lower than all algorithms.

Table 4.5: Logistic Regression

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Target** | **Precision** | **Recall** | **F1-Score** | **Support** |
| bangladesh | 0.73 | 0.54 | 0.62 | 290 |
| business | 0.86 | 0.88 | 0.87 | 328 |
| cricket | 0.96 | 0.95 | 0.95 | 284 |
| crime | 0.81 | 0.88 | 0.84 | 296 |
| education | 0.96 | 0.92 | 0.94 | 317 |
| entertainment | 0.92 | 0.97 | 0.94 | 288 |
| football | 0.96 | 0.97 | 0.97 | 305 |
| life | 0.85 | 0.84 | 0.84 | 298 |
| politics | 0.83 | 0.91 | 0.87 | 300 |
| science-tech | 0.88 | 0.93 | 0.91 | 294 |
|  |  |  |  |  |
| Accuracy |  |  | 0.88 | 2500 |
| Macro AVG | 0.88 | 0.88 | 0.88 | 2500 |
| Weighted AVG | 0.88 | 0.88 | 0.88 | 2500 |

**4.3.5: SGD classifier**

The precision recall SGD classifier and the f1 score matrix produce the best macro and weighted averages. The accuracy macro and weighted average are 0.91 and 0.91, respectively, as are the recalls. entertainment Aim for Accuracy has the highest score, and Recall receives the greatest score in entertainment, both at 0.98.

Table 4.6: SGD Classifier

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Target** | **Precision** | **Recall** | **F1-Score** | **Support** |
| bangladesh | 0.74 | 0.64 | 0.69 | 237 |
| business | 0.9 | 0.92 | 0.91 | 270 |
| cricket | 0.96 | 0.98 | 0.97 | 242 |
| crime | 0.85 | 0.88 | 0.86 | 249 |
| education | 0.95 | 0.96 | 0.96 | 263 |
| entertainment | 0.95 | 0.97 | 0.96 | 243 |
| football | 0.97 | 0.98 | 0.98 | 247 |
| life | 0.94 | 0.87 | 0.9 | 250 |
| politics | 0.89 | 0.92 | 0.9 | 251 |
| science-tech | 0.91 | 0.95 | 0.93 | 248 |
|  |  |  |  |  |
| Accuracy |  |  | 0.91 | 2500 |
| Macro AVG | 0.91 | 0.91 | 0.91 | 2500 |
| Weighted AVG | 0.91 | 0.91 | 0.91 | 2500 |

**CHAPTER 5**

**Conclusion and Future Work**

**5.1 Introduction**

In recent years, the press recommendations system has been widely employed to give better recommendations for users so that people may consume online news from multiple sources. The magazine's suggestion system contains a variety of variations, the majority of which are communicated via the news section. Among these problems are worries about speed, establishing reader preferences for dynamic news, the quality of material, and the consequences of press recommendations on user behavior. General suggestions are insufficient to generate new recommendations since they must be extensively updated, amended, or enlarged. Many of the shortcomings of traditional advice have lately been solved by learning and intense training methods. When we look at the Bengali news referral system, we discover that there is no such technology that operates online in Bangladesh. We are driven to do our job in order to fix this problem. In our investigation, we gathered information from several Bengali newspapers. Then we analyze this news.

**5.2 Conclusion**

In our obtained data, we created an artificial intelligence-based program. Our technique offers excellent outcomes while having minimal downsides. This is due to the fact that we cannot employ a really big data collection. We only utilize 10,000 data points in our work. Another barrier is that we exclusively operate in Bengali, although the press reports in English. We are unable to examine this sort of estimate. We will improve our work in the future by incorporating the following features.

**5.3 Implication for Future Study**

* We applied machine algorithms in our present task. We plan to use a deep learning algorithm to examine our data sets in the future.
* In the future, we will expand the amount of our data collection.
* There is no huge data frame since we employ sparks in our work; but, in the future, we will use a large data data frame to get more accurate findings.
* Our model does not support several languages. In the future, we will build our multilingual model by including various language data.
* Recent sophisticated algorithms, such as bert and distilbert, can be employed to better properly categorize our dataset.

**REFERENCE**

[1] Morteza Zihayat, Anteneh Ayanso, Xing Zhao, Heidar Davoudi, Aijun An, A utility-based news recommendation system, Decision Support Systems, Volume 117, 2019, Pages 14-27

[2] M. N. M. Adnan, M. R. Chowdury, I. Taz, T. Ahmed and R. M. Rahman, "Content based news recommendation system based on fuzzy logic," 2014 International Conference on Informatics, Electronics & Vision (ICIEV), 2014, pp. 1-6

[3] C. Feng, M. Khan, A. U. Rahman and A. Ahmad, "News Recommendation Systems - Accomplishments, Challenges & Future Directions," in IEEE Access, vol. 8, pp. 16702-16725, 2020

[4] Walter, F.E., Battiston, S. & Schweitzer, F. A model of a trust-based recommendation system on a social network. Auton Agent Multi-Agent Syst 16, 57–74 (2008).

[5] H. Tan, J. Guo and Y. Li, "E-learning Recommendation System," 2008 International Conference on Computer Science and Software Engineering, 2008, pp. 430-433

[6] Goyani, Mahesh; Chaurasiya, Neha. “A Review of Movie Recommendation System”. ELCVIA: electronic letters on computer vision and image analysis, [online], 2020, Vol. 19, Num. 3, pp. 18-37, https://raco.cat/index.php/ELCVIA/article/view/373942 [View: 3-02-2022]

[7] Gábor Takács, István Pilászy, Bottyán Németh, and Domonkos Tikk. 2007. Major components of the gravity recommendation system. SIGKDD Explor. Newsl. 9, 2 (December 2007), 80–83.

[8] Alejandro Montes-García, Jose María Álvarez-Rodríguez, Jose Emilio Labra-Gayo, Marcos Martínez-Merino, Towards a journalist-based news recommendation system: The Wesomender approach, Expert Systems with Applications, Volume 40, Issue 17, 2013, Pages 6735-67419o

[9] Raza, S., Ding, C. News recommender system: a review of recent progress, challenges, and opportunities. Artif Intell Rev 55, 749–800 (2022)

[10] M. H. Rahman, M. S. Islam, M. M. U. Jowel, M. M. Hasan and M. S. Latif, "Classification of Book Review Sentiment in Bangla Language Using NLP, Machine Learning and LSTM," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-5, doi: 10.1109/ICCCNT51525.2021.9580116.

[11] Sri Krishnan, 6 - Machine learning for biomedical signal analysis, Editor(s): Sri Krishnan, Biomedical Signal Analysis for Connected Healthcare, Academic Press, 2021, Pages 223-264

[12] Sarica A, Cerasa A and Quattrone A (2017) Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Front. Aging Neurosci. 9:329.

**APPENDIX**

The first hassle we had while doing the evaluation changed into organizing the analytical method for our investigation. It wasn`t trendy job, and little have been accomplished on this problem previously. As a result, we were not capable of get a whole lot assist from any source. We additionally began out collecting records through hand. After a prolonged time of difficult labor, we is probably capable of acquire it.

**PLAGIARISM REPORT**

