

**NewsML: A Machine Learning Approach for Bangla News
Classification**

**BY
Abid Hasan Rafi**

ID: 201-16-495

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Computing & Information System

Supervised By
Sonia Nasrin
Lecturer
Computing & Information System
Daffodil International University



**DEPARTMENT OF COMPUTING AND INFORMATION SYSTEM
DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH**

Submission date 23 September 2024

APPROVAL

This thesis titled “NewsML: A Machine Learning Approach for Bangla News Classification”, submitted by Abid Hasan Rafi, ID No: 201-16-495 to the Department of Computing & Information Systems, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computing & Information Systems and approved as to its style and contents. The presentation has been held on- 23-09-2024.

BOARD OF EXAMINERS



Mr. Md. Sarwar Hossain Mollah
Associate Professor and Head
Department of Computing & Information System
Faculty of Science & Information Technology
Daffodil International University

Chairman



Md. Nasimul Kader
Assistant Professor
Department of Computing & Information System
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Mr. Mehedi Hasan
Lecturer
Department of Computing & Information System
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Saifuddin Md. Tareeq
Professor
Department of Computer Science and Engineering
University of Dhaka, Dhaka

External Examiner

DECLARATION

I hereby declare that; the work in this dissertation titled “**NewsML: A Machine Learning Approach for Bangla News Classification**” has been done by me under supervision of **Sonia Nasrin** Lecturer, department of Computing and Information System (CIS) of Daffodil International University. I am also declaring that this project or any part of there has never been submitted anywhere else for the award of any educational degree like, B.Sc., M.Sc., Diploma or other qualifications.

Supervised By

Sonia Nasrin
23/07/24

Sonia Nasrin

Lecturer

Department of CIS

Daffodil International University

Submitted By

Abid Hasan Rafi

Abidn Hasan Rafi

ID: 201-16-495

Department of CIS

Daffodil International University

ACKNOWLEDGEMENT

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

I am grateful and wish our profound indebtedness to **Sonia Nasrin, Lecturer**, Department of Computing and Information System (CIS) Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of *Artificial Intelligence* 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 stages have made it possible to complete this project.

I would like to express our heartiest gratitude to the **Mr. Sarwar Hossain Mollah, Head of the Department of Computing and Information System (CIS)**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computing and Information System (CIS), Daffodil International University.

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

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

ABSTRACT

The importance of effectively categorizing news information is rising with the expansion of digital media in Bangladesh. The problem of sorting NewsML headlines into six individual categories Entertainment, Life, Business, Education, World, and Sports is addressed in this work. The goal is to determine how well six machine learning algorithms—Multi, KNN, DT, RF, LR, and SVC—perform in automated classification. The study aims to improve the performance of the digital news platform user experience and content management system by developing a solid framework for Bangla news classification. The dataset comprises five thousand headlines from prominent NewsML portals, including Prothom Alo, Kaler Kontho, BDNews, and News24. The research shows that the SVC algorithm obtains the best accuracy of 0.99% on a 20% test dataset after extensive preprocessing and analysis. This finding demonstrates how well SVC works when classifying news headlines and dealing with the subtleties of Bangla text.

TABLE OF CONTENTS

CONTENTS	PAGE NO.
Declaration	i
Abstract	ii
List of Tables	v
List of Figures	vi
CHAPTER	
CHAPTER 1: INTRODUCTION	1-4
1.1 Overview	1
1.2 Research Objective	2
1.3 Motivation	2
1.4 Rational of Study	3
1.5 Research Question	3
1.6 Project Outcome	4
CHAPTER 2: BACKGROUND STUDY	5-9
2.1 Introduction	5
2.2 Related Work	5
2.3 Perspective of Bangladesh	8
2.4 Scope of this problem	8
2.5 Challenge of this work	9
CHAPTER 3: Methodology	10-14
3.1 Overview	10
3.2 Data Collection	11
3.3 Data Preprocessing	12
3.4 Classification	13
3.5 Implementation Requirement	14

CHAPTER 4: RESULT ANALYSIS	15-22
4.1 Experimental Result	15
5.2 Classification report	15
5.4 Decision	21
5.5 Evaluation	22
CHAPTER 5: CONCLUSION	23
CHAPTER 6: FUTURE WORK	24
REFERENCES	25
PLAGIARISM REPORT	26

LIST OF TABLES	
Table	PAGE NO.
Figure 1 Tokenization table	12
Figure 2 Accuracy Table	15
Figure 3 SVC Classification Report	15
Figure 4 RF Classification Report	17
Figure 5 Multi Classification Report	18
Figure 6 LR Classification Report	19
Figure 7 KNN Classification Report	20
Figure 8 DT Classification Report	21

LIST OF FIGURES	
FIGURES	PAGE NO.
Figure 3.1 Mythology Diagram	10
Figure 3.2 Representation of Imbalanced dataset	13
Figure 3.3 Representation of Imbalanced dataset	13
Figure 3.3 Evaluation graph	22

LIST OF ACRONYMS

SVC Support Vector Classifier

RF Random Forest

NN Neural Networks

DT Decision Trees

KNN K-Nearest Neighbors

LR Logistic Regression

CHAPTER 1

INTRODUCTION

1.1 Overview

An increasing number of individuals are presently highly engaged on a diverse range of social media platforms, such as Facebook, YouTube, Twitter, and several more. The present investigation centers on analyzing NewsML and its intricate network of interconnections. With the rapid expansion of digital information, particularly in languages other than English, such as Bangla, it is crucial to have automated classification and processing of online news. More than 250 million people in Bangladesh and certain parts of India speak Bengali as their primary language. More attention should be given to the need to research Bangla-specific NLP algorithms, particularly for text categorization. Considering the extensive volume of news published daily, implementing an automated system capable of classifying news articles into pertinent categories will significantly enhance retrieving and managing information. The Bangla language's intricate alphabet and expansive morphology pose challenges for automating headline-based news categorization. The absence of annotated datasets for Bengali text categorization poses challenges in developing effective machine-learning models. To address these obstacles, this study gathers 5,000 Bangla news headlines from reputable sources such as Prothom Alo, Kaler Kontho, BDNews, and News24, covering various topics, including Entertainment, Life, Business, Education, World, and Sports. Efficiently categorizing news items enhances the structural organization of material and improves the user experience for Bangla-speaking consumers. News categorization algorithms utilized by automated systems assist news aggregators, search engines, and content recommendation systems. These approaches enable them to provide users with far more pertinent content. The development of Bangla NLP is comparatively less advanced than English NLP, and this study contributes to that knowledge. To conduct NLP research in underrepresented languages and improve digital material accessibility, it is imperative to develop robust Bangla text classification models. In this work, the SVM model performs more than other machine learning algorithms in categories of Bangla news headlines. Support Vector

Machines (SVM) are well-suited for text classification tasks involving high-dimensional data and sparse matrices, making them particularly suitable for Bangla news classification. The present study aims to facilitate Bangla natural language processing research and contribute to the development of automated news classification systems. Potential future studies can utilize the results to develop more sophisticated models, such as those based on deep learning, to enhance the classification of Bangla texts.

1.2 Research Objective

- Design and implement a robust machine learning system capable of classifying Bangla news headlines.
- Aggregate and accurately categorize a total of five thousand news articles in the Bangla language.
- This paper aims to compare and contrast six machine learning methods.
- Conduct an accuracy evaluation of the model by employing dataset splits of 25%, 25%, 30%, and 35% in order to reveal trends associated with data adequacy.
- Ascertain the optimal algorithm for the classification of Bangla news.
- Design a Bangla natural language processing system by researching machine learning methods for text categorization in the Bangla language.

1.3 Motivation

- As Bangla digital news volume increases, it is necessary to enhance information management to facilitate automated categorization.
- In contrast to English, the Bangla language needs more NLP resources despite its widespread use.
- Applying machine learning methods for classifying Bangla news headlines is necessary to enhance content retrieval.
- Implementing automated news categorization would enhance the user experience of news websites by increasing the accessibility of pertinent information.

- This work explores and enhances machine learning methods for text categorization to address a need for more expertise in natural language processing for the Bangla language.
- Adding to Bangla NLP can accelerate the development of language technology for poorly represented languages.

1.4 Rational of Study

The increasing prevalence of digital news necessitates the automation of Bangla news categorization to enhance accessibility and management. The predominant emphasis of natural language processing (NLP) research and tools is on English and other extensively spoken languages, which poses challenges for developing Bangla resources. This study uses machine learning techniques to detect Bangla news articles based on their headlines, a domain that has received less scholarly attention. This research justifies the need for efficient classes of news articles in Entertainment, Life, Business, Education, World, and Sports. This category will facilitate the efficient retrieval of pertinent items for news site users. Expanding the internet-based Bangla-speaking community will elevate the requirement for automated content management solutions, underscoring the necessity for natural language processing systems tailored to the Bangla language. The results will address a research gap in Bangla natural language processing and contribute to the progress of multilingual natural language processing.

1.5 Research Question

- Which machine learning approach achieves the highest performance when categorizing Bangla news headlines?
- What is the correlation between the accuracy of KNN, SVC, Multinomial Naive Bayes, RM, DT, and SVM?
- To what extent do tokenization, feature extraction, and stop word removal impact the preparation of Bangla news categorization machine learning models?

- The classification of Bangla news dataset categories poses significant challenges; which algorithms provide superior performance in this task?
- What impact will the results of this study have on Bangla NLP and automated content management systems?

1.6 Expected Outcome

- Choosing the optimal machine learning technique for Bangla news headline classification, with SVMs predicted to perform better.
- This paper examines KNN, SVC, Multinomial Naive Bayes, RM, DT, and SVM pros and cons. The results show its scalability and resilience over dataset divisions.
- I advocate better Bangla NLP text preparation and examine how preprocessing affects model accuracy.
- A look at the challenges of classifying some news stories to improve models in these complex domains.
- By thoroughly examining machine learning methodologies, this study advances Bangla NLP research. It prepares automated Bangla text categorization for future enhancements.

CHAPTER 2

BACKGROUND STUDY

2.1 Introduction

This chapter discusses news article categorization, focusing on Bangla text. First, it examines NLP and electronic text categorization. Then, it covers K-Nearest Neighbours (KNN), Support Vector Machines (SVM), Multinomial Naive Bayes, Regression Methods (RM), and Decision Trees for text categorization. It then examines Bangla NLP's challenges and advancements, focusing on Bangla's distinctive linguistic traits that affect text processing and categorization. The chapter also reviews prior studies and methods for categorizing news pieces, identifying gaps in present research, and the need for a Bangla-language study. This chapter presents a complete background to help understand the research problem, justify the investigation, and define the research goals in later chapters.

2.2 Related work

Haque et al. [1] have created a machine-learning approach for classifying Bangla news. Categorizing a large number of Bangla news articles is a time-consuming task; this study intends to streamline it. The suggested model utilizes algorithmic techniques such as NB, DT, and RM to classify Bangla news stories into many categories, such as sports, politics, entertainment, and others. The Random Forest technique achieved the highest success rate of 92.83 percent among the methods.

Hossain et al. [2] introduce a statistical model for categorizing news stories in the Bangla language. This approach prioritizes data retrieval and analysis by employing Bangla classification. A machine learning model informs the design of the bag-of-words representation, and TF-IDF used to represent news items. Feature-based classification of news items is performed using NB, SVM, and RM Machine algorithms. Indications of the paper's correctness are promising.

Roy et al. [3] conducted a machine learning-based classification of Bangla news articles. This project aims to employ their approach for efficient, systematic organization and retrieval of Bangla news material. Empirical evidence has consistently demonstrated that

the SVM algorithm consistently yields the highest levels of accuracy. This paper establishes that machine learning algorithms can identify Bangla news, thus enhancing text classification and natural language processing.

Rahman et al. [4] evaluate the classification of Bangla news items by employing neural networks on two datasets. In Dataset I, Prothom Alo articles are utilized, whereas Dataset II relies on Kaggle data. Performance optimization is achieved by custom stop-word preprocessing. NB, SVM, and Multi-layer Neural Networks perform better than other methods. SVM, logistic models, and multi-layer dense neural networks achieved accuracy rates of 94.51%, 94.60%, and 95.50% respectively.

A. N. Chy et al. [5] propose the automated categorization of Bangla news articles to enhance information organization and retrieval. Their approach is based on the NB, a reliable text classification technique. Given the results, it is advisable to employ stemming methods to reduce the number of dimensions in the feature space, tokenize news items, and eliminate stop words. The authors comprehensively explain their methodology for implementing and assessing the NB.

Chowdhury et al. [6] present a multifunction model consisting of CNN, LSTM, and GloVe vectorization to classify Bangla news. GloVe pre-trained word embeddings represent Bangla text with semantic encoding. Extracting key features from vectorized representations is achieved by employing LSTM and CNN architectures. This effort aims to tokenize Bangla news and statistically express it using GloVe embeddings. The integration of LSTM and CNN layers enables the extraction of sequential and local context from news items. This paper provides a comprehensive analysis of the experimental accuracy of the leading algorithm. The suggested system achieves an 87% accuracy in categorizing Bangla news, while the exact value is unknown. Optimise cat learning by incorporating CNN, LSTM, and GloVe vectorization.

The reliability of classified Bangla articles has been established by M. Tanvir Alam et al. [7]. The authors' objective was to classify Bengali written content accurately. For this objective, the Bangla article categorization dataset BARD was utilized. Researchers employed a range of methodologies to determine the most effective approach. Information of significant value was extracted from the preprocessed dataset using natural language

processing and feature selection. A comparative analysis was conducted on k-NN, RM, and SVMs to determine the optimal model. This work utilized these algorithms and the BARD dataset. Following the effective implementation of the strategy, the leading algorithm successfully detected Bangla articles.

In their study, Rahman et al. propose a methodology for categorizing news articles that are reported in the Bangla language. The author developed a Graph Convolutional Network (GCN) approach to categorize Bengali news articles optimally. Graph convolutional networks are well-suited for classifying graph-structured data due to their ability to understand and exploit complex structural connections and interdependencies. To substantiate their hypothesis, the researchers carried out systematic experiments. Computational neural networks (GCNs) were used to extract relevant information from graph models that included Bangla news articles. In GCN-based systems, the semantic information in textual material and its connections to other articles are utilized to classify content precisely.

The study conducted by N. Humaira et al. [9] extensively examines the application of deep learning and machine learning methods in the classification of Bengali documents. Consequently, the writers devised an algorithm to categorize news articles composed in Bengali. An investigation was conducted on extracting multiple features, deep learning, and machine learning techniques. An analysis was performed using TF-IDF, Word2Vec, and Doc2Vec to identify significant characteristics in Bengali news articles. The approach utilized CNN, RM, and SVM classifiers. The experimental study confirms the effectiveness of the proposed models. The highest-performing system attained a 97% accuracy rate in classifying Bengali news documents using feature extraction and deep learning techniques.

Adrita et al. [10] establish a dataset and assess machine learning methods for classifying sports news items. The authors have developed a dataset suitable for multi-class classification to enhance the organization of sports news material. Researchers experiment with several machine-learning algorithms. The models are trained using the recently created dataset and the following algorithms: k-NN, NB, DT, Reinforcement Learning, and SVM. The precision, accuracy, recall, and F1 score determine the efficacy of a machine-

learning approach. The empirical evidence demonstrates the accuracy of the models. An F1-Score of 97.06 indicates that the SVM system attained the maximum level of accuracy in categorizing sports news pieces.

Rahman et al. [11] presented a classification model for Bangla language documents. They aimed to accurately categorize the Bangla language using Recurrent Neural Networks (RNNs). To extract contextual information from sequential input effectively, the researchers utilized BiLSTM architectures. The model acquires knowledge of the connections between Bangla papers via forward and backward data flow processes. Presumably, the authors evaluated their model using accuracy, precision, recall, or F1-score, as the analysis indicates that the most effective approach achieves an accuracy of 98.33%. These metrics would indicate the classification performance of BiLSTM on Bangla documents.

2.3 Perspective of Bangladesh

Bangladesh's fast-growing digital media and Bangla-speaking internet users make automated news classification systems crucial. Bangladesh digitizes swiftly, creating online news and information. Due to this rapid expansion, content management and categorization are tougher, making it harder for users to find important information quickly. Information retrieval and user experience would improve if Bangla news headlines were automatically classified into Entertainment, Life, Business, Education, World, and Sports. Thus, news platforms may provide more relevant and tailored content, boosting reader satisfaction. Bangladesh loves cricket, and appropriate classification can help fans discover news and updates. Bangla natural language processing resources, which are behind English, are improved by this work. This project creates Bangla-language and culture-specific tools using new machine learning. It also helps local tech.

2.3 Scope this problem

Machine learning approaches for the classification of Bangla news headlines need to be developed and tested to solve this problem. It is crucial to preprocess a massive Bangla news dataset from News24, Prothom Alo, Kaler Kontho, and BDNews. In order to train a machine learning model, the input text must first be tokenized, stop words removed, and

features extracted. We test the accuracy of K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Multinomial Naive Bayes (RM), Decision Tree (DT), and Support Vector Machine (SVM) in classifying news headlines into Entertainment, Life, Business, Education, World, and Sports across 20%, 25%, 30%, and 35% dataset splits. The impact of script quirks and linguistic complexities on the performance of Bangla text categorization models is examined in this review. At last, the research delves into the ways in which the expanding digital media ecosystem in Bangladesh could benefit the user experience, information retrieval, and content management of the country's news platforms through enhanced news classification systems.

2.4 Challenges of this work

It is challenging to categorize Bangla news headlines. The intricate script and grammar of Bangla make tokenization and stop-word removal challenging tasks in text processing. Complexity can affect the accuracy of preprocessing, which is crucial for machine learning models. Bangla news headline datasets that are accurate and balanced across multiple sources are rare. If we want to build good classification models, we need a dataset that includes Entertainment, Life, Business, Education, World, and Sports. It becomes more complicated when preprocessing and feature extraction are adjusted to fit the Bangla dialect. Because dataset splits and text quality impact performance, testing, and analysis are necessary to discover the best machine-learning models. Issues with underrepresented news categories, or category imbalance, could compromise classification accuracy and call for remedies. To apply the research findings to real-world systems, successful classification algorithms need to be integrated into dynamic news platforms and adapted to changing news material. All of these problems show how important it is to build and test a Bangla news classification system thoroughly.

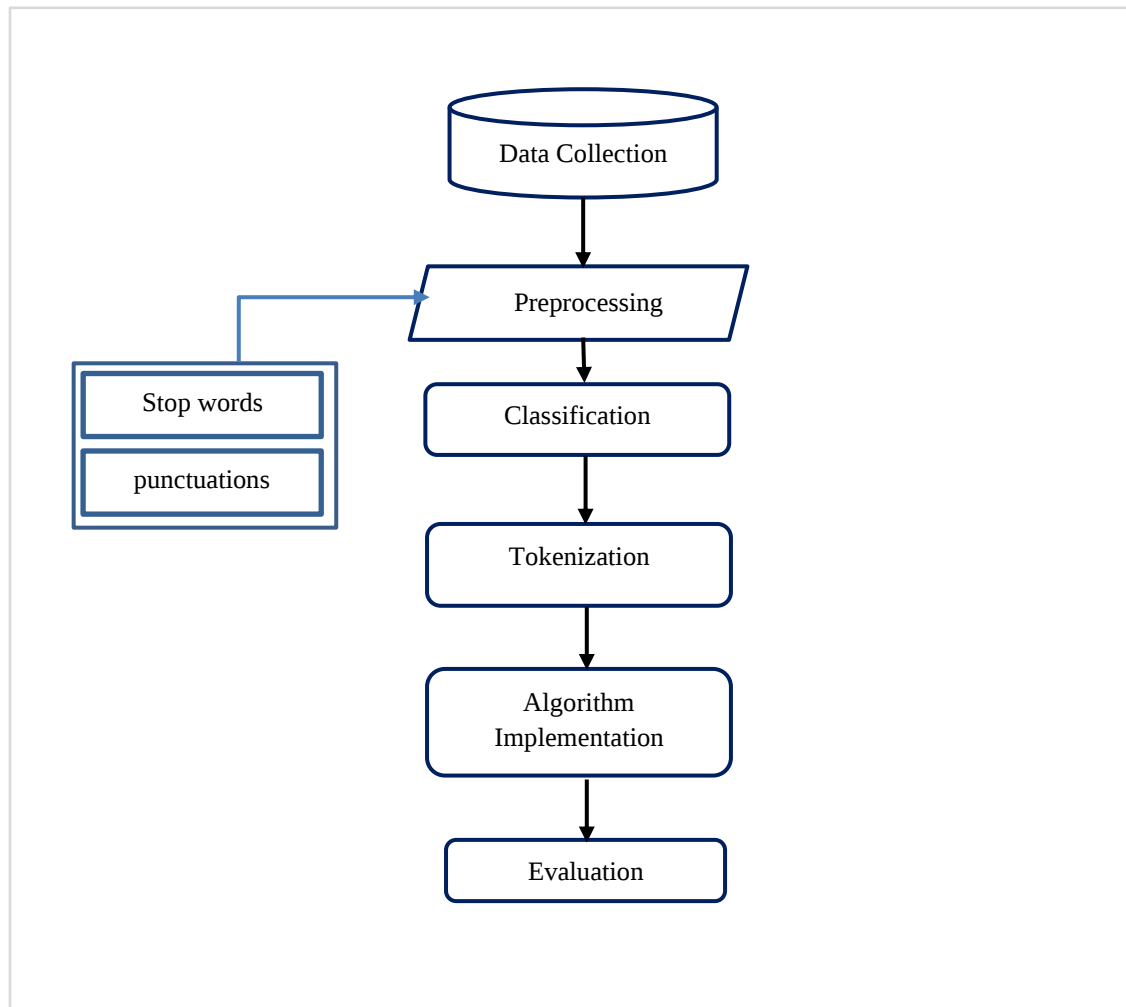
CHAPTER 3

METHODOLOGY

3.1 Overview

A "research technique" is a systematic approach used to define, choose, organize, and evaluate pertinent facts and derive conclusions from this data. The methodology portion of a research paper enables readers to assess the credibility and coherence of the study. The study in section 3.2 is conducted through six sequential stages. The procedures are as follows.

Figure 3.1 Methodology Diagram



3.2 Data collection

This study aimed to test various machine learning approaches for text classification using Data scraping and a massive dataset of Bangla news headlines. Prothom Alo, Kaler Kontho, BDNews, and News24 are the four main Bangla news portals from which we culled 5,000 news items. In selecting these sources, we aimed for a variety of article genres and news substance. The following news categories were compiled: Entertainment, Life, Business, Education, World, and Sports. This classification ensures that the dataset has a wide range of subjects covered, which is useful for training and testing classification models across domains. To ensure data quality and relevance, headlines were gathered according to recency and correctness. Data Preprocessing included removing extraneous information, handling special characters, and standardizing the text for analysis after data collection.

3.4 Data Preprocessing

Machine learning analysis of Bangla news headlines requires preprocessing. To make the text more readable and consistent, text cleaning involves removing extraneous punctuation, special characters, and HTML elements. Tokenization uses the intricacies of the Bangla script to divide the text into words or phrases, allowing for precise segmentation. Stop words are eliminated to get rid of popular, low-information keywords that don't contribute to classification. To eliminate these, Bangla employs a stop word list that is unique to the language. Words are reduced to their most basic forms through stemming and lemmatization, which merge variants into one. The next step, feature extraction, employs TF-IDF to convert the text into numerical forms representing each phrase's importance to the dataset. Last, data normalization ensures that everyone is equally likely to be classified by scaling attributes to a specific range. To train and evaluate machine learning models, these preprocessing processes organize Bangla news headlines.

Tokenization

Tokenization is a preprocessing approach that natural language processing (NLP) uses to break down big Bangla news headlines into smaller, more manageable pieces. Cutting out unnecessary spaces and punctuation marks is the first step in deconstructing the text. [12] Words, phrases, or symbols can make up these components. The complexity of the Bangla

script, with its unique characters and conjuncts, makes it difficult for tokenization to detect word borders accurately. As part of tokenization, spelling variations are standardized, and special cases that may require processing are handled, such as dates, numbers, and named objects. Successful machine learning and classification efforts are built upon the efficient extraction and analysis of features made possible by tokenizing the raw text.

TABLE 1 TOKENIZATION TABLE

Raw Data	Type	Tokenized data
১৯ আগস্ট বিশ্ব আলোকচিত্র দিবস উদযাপন উপলক্ষে দিনব্যাপী ফটোবুক প্রদর্শনী আয়োজন	Life	'১৯', 'আগস্ট', 'বিশ্ব', 'উদযাপন', 'আলোকচিত্র', 'উপলক্ষে', 'দিনব্যাপী', 'ফটোবুক', 'প্রদর্শনী', 'আয়োজন',
মাম্মা দেব বড় ভাইয়ের ছেলে তড়িৎ দেব বিরুদ্ধে	Entertainment	'মাম্মা', 'দেব', 'বড়', 'ছেলে', 'তড়িৎ', 'দেব', 'বিরুদ্ধে',
ক্রিয়ার কালপ্রিয় শিক্ষার্থীরা আজ তোমাদের জন্য	Education	'ক্রিয়ার', 'কালপ্রিয়', 'গত', 'শিক্ষার্থীরা', 'আজ', 'তোমাদের', 'জন্য',
কম্পিউটার সমিতি বিসিএস হেলথ অ্যান্ড হাইজেনিক শ্রেণিতে সুপ্রিম কোর্টের জ্যেষ্ঠ আইনজীবী রফিকউল	Business	'কম্পিউটার', 'সমিতি', 'বিসিএস', 'হেলথ', 'অ্যান্ড', 'হাইজেনিক', 'শ্রেণিতে', 'সুপ্রিম', 'কোর্টের', 'জ্যেষ্ঠ', 'আইনজীবী', 'রফিকউল',
পানামা খাল থেকে আটক উত্তর কোরীয় জাহাজে পাওয়া	World	'পানামা', 'খাল', 'লাশ', 'থেকে', 'আটক', 'উত্তর', 'কোরীয়', 'জাহাজে', 'পাওয়া',
মেসিনেইমার দলের অংশ দলের দৃষ্টি প্রতিভাবানের ওপর নিবন্ধ অবচেতনেই খেলোয়াড়েরা	Sports	'মেসিনেইমার', 'দলের', 'অংশ', 'দলের', 'দৃষ্টি', 'প্রতিভাবানের', 'ওপর', 'নিবন্ধ', 'অবচেতনেই', 'খেলোয়াড়েরা',

3.4 Classification

Figure 3.2 and 3.3 shows the imbalanced and balanced of the dataset. This figure shows the six labels of my project: cricket, crime, politics, business, life, and football. Before classification of data set it seen to be there was different count of labels. It was very ineffective dataset. After classification, this project data seems well balanced. Label Education, Entertainment, Business, life, World, Sports show 16.7%. No far values are

seen, all of labels constancy is quals. So, it is said that this data set is a good dataset for this project.

Figure 3.3 Representation of Imbalance data set

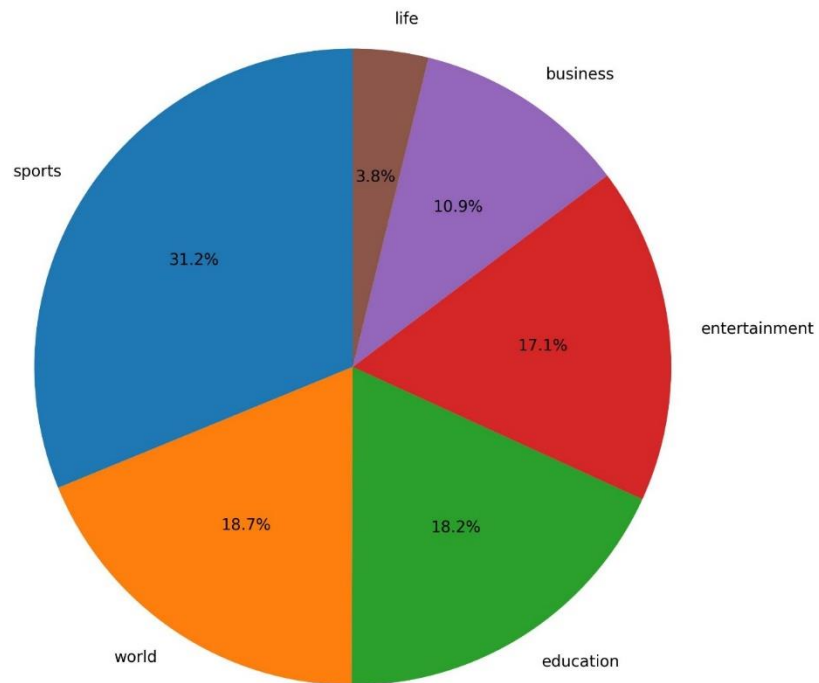
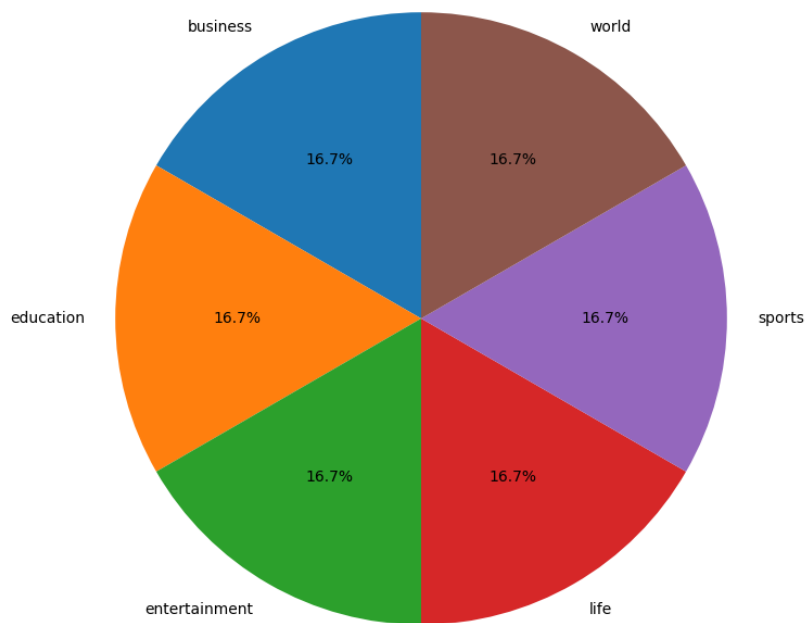


Figure 3.3 Representation of balance data set



3.5 Implementation requirement

To categorize Bangla news headlines, several criteria must be met. Initially, it is essential to establish a robust data collection system, including a dependable storage solution and web scraping or application programming interfaces (APIs) to extract headlines from news sources such as Prothom Alo, Kaler Kontho, BDNews, and News24. [13] Additionally, necessary preparatory methods involve scripts for text cleaning, tokenization based on Bangla norms, and elimination of stop words. Python, Scikit-learn, or TensorFlow are required for constructing and assessing models using KNN, SVC, and SVM in machine learning. Efficient feature extraction techniques, such as TF-IDF, are essential for transforming text into numerical data suitable for analysis. Integrating model performance metrics such as accuracy and F1-score is a crucial requirement. Considerable memory and computational power are required for data processing and model training. Finally, if feasible, enhancing the user interface of the classification system will further improve its accessibility and usability. To construct a dependable database for classifying Bangla news headlines, it is necessary to satisfy these requirements.

CHAPTER 4

Result Analysis

4.1 Overview

Two of the six research strategies included in the study were preprocessed data analysis and general research. I compiled the accuracy metric for all the approaches in Table 2 to expedite the comparison of algorithms.[14] Complete accuracy data are available in this table. An Accuracy investigation was conducted using sample sizes of 25%, 30%, 35%, and 20%. An analysis of these six research methodologies revealed unexpected results. Based on my investigation, all the following algorithms have proven effective: KNN, Multi, RM, DT, and SVC. SVC achieved a 90% accuracy on a 20% test dataset. This run was the most successful execution of the machine up until that point.

TABLE 2 ACCURACY TABLE

	Algorithm					
	RM	Multi	SVC	KNN	LR	DT
Accuracy	0.99%	0.91%	0.99%	0.93%	0.97%	0.96%

4.2 Classification Report

Classification reports are commonly used to assess the effectiveness of machine learning algorithms. A trained classification model's performance and reliability can be evaluated using recall, accuracy, F1 Score, and support metrics. The SVC categorization is extensively utilized in both academic and corporate settings. For my project, I've settled on sixteen tiers of widely-used content. The SVC classification method is a powerful machine-learning tool that ranks data points according to their feature vector.

SVC

Table 3 displays the classification table for SVC methods. At the Business level, this algorithm achieved a precision rate of 1.00%. A precision of 1.00% is shown by sports and

world. Education achieved a recall rate of 1.00%. Life rates precision at 0.99%. Entertainment achieved a 0.99% average for three matrices. The remaining two micro and Wright averages provide a 0.99% accuracy rate for all three matrices. The accuracy of this method was 0.95% when the test data consisted of 1380 rows. This algorithm demonstrates excellent performance in accurately detecting the six labels.

TABLE 3 SVC CLASSIFICATION REPORT

label	SVC			support
	Precision	Recall	F1-score	
Business	1.00	1.00	1.00	239
Education	0.99	0.99	0.99	233
Entertainment	0.99	0.98	0.99	226
life	0.99	1.00	0.99	212
Sports	1.00	1.00	1.00	226
World	1.00	0.99	1.00	224
accuracy	0.99			1380
Macro Avg	0.99	0.99	0.99	1380
Weighted avg	0.99	0.99	0.99	1380

RF

Table 4 displays the categorization table for Random Forest (RF) algorithms. At the business, Education, Entertainment labels, this method achieved a precision rate of 1.00%. Sports records a precision rate of 0.99%. World achieved a recall rate of 0.99%. Life displays a precision of 0.99%. The remaining two micro and Wright averages consistently indicate a value of 0.99% for all three matrices. The accuracy of this method was 0.99%,

and the test data consisted of 6277 instances. The score achieved by this algorithm was quite low.

TABLE 4 RM CLASSIFICATION REPORT

label	RM			support
	Precision	Recall	F1-score	
Business	1.00	1.00	1.00	239
Education	1.00	0.99	1.00	233
Entertainment	1.00	0.98	0.99	226
life	0.99	1.00	1.00	212
Sports	0.99	0.99	0.98	226
World	0.97	0.99	0.99	224
accuracy	0.99			1380
Macro Avg	0.99	0.99	0.99	1380
Weighted avg	0.99	0.99	0.99	1380

Multi

Table 5 displays the categorization table for multi-algorithms. The recall rate achieved by this algorithm in the business setting was 0.99%. The precision recorded in Education is 1.00%. Entertainment had a precision rate of 0.97%. The Sports category has a recall rating of 0.97%. Live displays a precision of 0.94%. World achieved a recall rate of 0.95%. The remaining two micro and Wright averages consistently indicate a value of 0.92% for all

three matrices. The accuracy of this method was 0.92% when the test data consisted of 1380 rows. The score achieved by this algorithm was relatively low.

TABLE 5 MULTI-CLASSIFICATION REPORT

label	Multi			support
	Precision	Recall	F1-score	
Business	0.90	0.99	0.94	239
Education	1.00	0.69	0.82	233
Entertainment	0.97	0.90	0.93	226
life	0.76	1.00	0.86	212
Sports	0.98	0.97	0.98	226
World	0.95	0.95	0.95	224
accuracy	0.92			1380
Macro Avg	0.92	0.92	0.92	1380
Weighted avg	0.92	0.92	0.92	1380

LR

Table 6 displays the classification table for LR algorithms. The algorithm achieved a precision rate of 1.00% in the business setting. Education demonstrates a precision rate of 0.97%. Entertainment achieved a precision rate of 0.99%. Sports has a precision rate of 1.00. Life exhibits a precision of 0.93%. A 0.98% increase was observed in world across three matrices. The remaining two micro and Wright averages consistently indicate a value of 0.92% for all three matrices. An accuracy of 0.98% was attained by this method on a test dataset of 6277.

TABLE 6 LR CLASSIFICATION REPORT

label	LR			support
	Precision	Recall	F1-score	
Business	1.00	0.99	0.99	239
Education	0.97	0.95	0.96	233
Entertainment	0.99	0.97	0.98	226
life	0.93	1.00	0.96	212
Sports	1.00	0.99	0.99	226
World	0.99	0.97	0.98	224
accuracy	0.98			1380
Macro Avg	0.98	0.98	0.98	1380
Weighted avg	0.98	0.98	0.98	1380

KNN

Table 7 displays the classification table for KNN algorithms. At the business level, this method achieved a precision rate of 0.96%. Education demonstrates a precision rate of 0.92%. Entertainment recorded a precision rate of 0.95%. Sports achieved a precision rate of 0.99%. Life exhibits a recall of 1.00%. A 0.97% increase was observed in world across three matrices. The remaining two micro and Wright averages indicate a statistical significance of 0.89% for all three matrices. The method attained an accuracy of 0.94% when the test data consisted of 6277 rows. The performance attained by this method was really subpar.

TABLE 7 KNN CLASSIFICATION REPORT

label	KNN			support
	Precision	Recall	F1-score	
Business	0.96	1.00	0.98	239
Education	0.92	0.91	0.92	233
Entertainment	0.95	0.92	0.94	226
life	0.84	1.00	0.92	212
Sports	0.99	0.91	0.94	226
World	0.97	0.98	0.92	224
accuracy	0.94			1380
Macro Avg	0.94	0.94	0.94	1380
Weighted avg	0.94	0.94	0.94	1380

DT

Table 8 presents the categorization table for decision tree algorithms. At the Business level, this algorithm achieved a precision rate of 0.98%. Sports demonstrates a precision rate of 0.96%. Education had a precision rate of 0.96%. The recall index for World is 0.97%. Life rates precision at 0.97%. A 0.94% increase was observed in Entertainment across three matrices. The other two micro and Wright averages indicate a mean value of 0.96% for all three matrices. The accuracy of this method was 0.96% when the test data consisted of 1380 rows.

TABLE 7 DT CLASSIFICATION REPORT

label	DT			support
	Precision	Recall	F1-score	
Business	0.98	0.99	0.98	239
Education	0.96	0.96	0.96	233
Entertainment	0.94	0.96	0.95	226
life	0.94	1.00	0.95	212
Sports	0.96	0.89	0.92	226
World	0.97	0.95	0.96	224
accuracy	0.96			1380
Macro Avg	0.96	0.96	0.96	1380
Weighted avg	0.96	0.96	0.96	1380

4.3 Decision

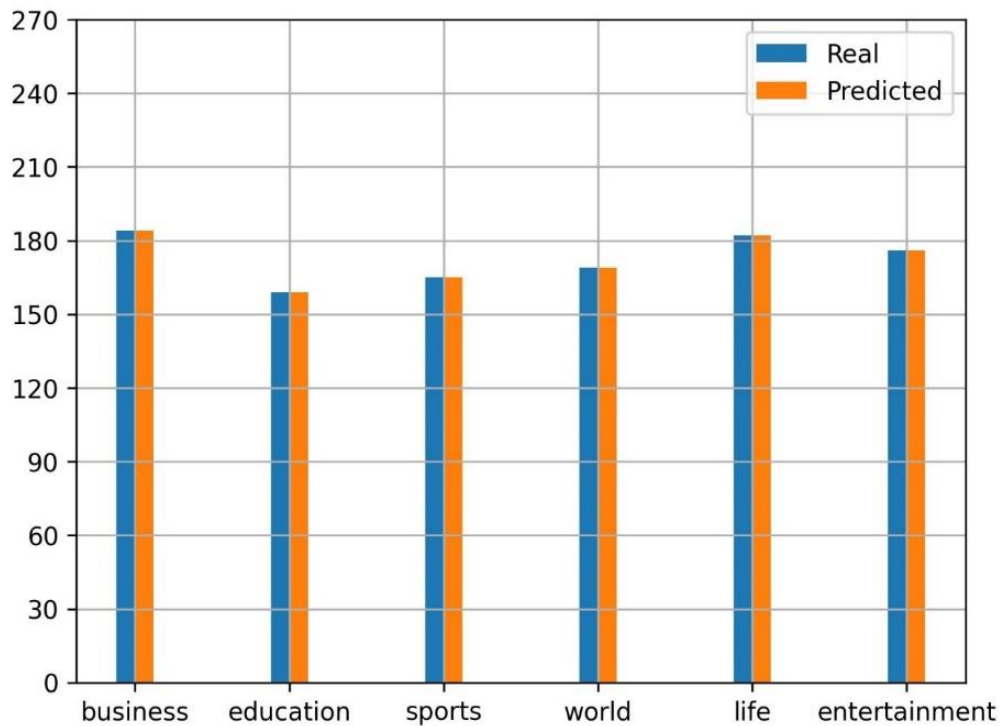
This paper presents the findings of research on the classification of Bangla news headlines. This paper assesses each of the six machine-learning approaches. Stochastic Vector Machines (SVC) achieved a 0.99% accuracy rate on the 20% test dataset, surpassing other methods. The primary focus of SVC is to classify Bangla news headlines into the following categories: Entertainment, Life, Business, Education, World, and Sports. Based on careful evaluation of all factors, the ultimate implementation of the Bangla news categorization system will use the SVC algorithm. [15] The SVM model has the highest classification reliability with an accuracy of 0.99%. Furthermore, it is possible to enhance and refine the SVM model in order to further enhance its performance. [16] The SVC method was selected based on its proven effectiveness, and additional refinement is being investigated

to guarantee optimal accuracy in practical applications. Such a result validates the objective of accurately classifying Bangla news headlines.

4.4 Evaluation

Figure 4.1 shows the evaluation graph for this . In this graph, blue and orange bars refer to the real vs. predicted data for this project. The selected algorithm can easily detect all six labels. There are a few mistakes here, but they are not huge. Overall, I scored well. As an observation of this graph, it can be said that SVC does very well for this sentiment dataset.

Figure 4.1 Evaluation graph



CHAPTER 5

Conclusion

This chapter will provide a concise overview of the main characteristics of the study and derive conclusions on the application of machine learning for the classification of Bangla news headlines. Entertainment, Life, Business, Education, World, and Sports were among the classifications employed in this study to segment news headlines. A comparative analysis was conducted on six distinct algorithms. The methods included were DT, NB, KNN, LR, Multi, and SVC. The SVC algorithm attained the maximum accuracy of 0.99% when applied to a 20% test dataset. This discovery provides further evidence of the higher performance of the SVC algorithm in comparison to other algorithms for Bangla news categorization. The efficacy of the SVC model demonstrates its ability to effectively manage the grammatical complexities of the Bangla language and accurately categorize a wide range of news topics. The study tackled several issues, including data quality, preprocessing protocols' efficacy, and the intricacy of the Bangla script. By successfully addressing these questions, the research significantly extends our understanding of the optimal preprocessing and machine-learning methods for Bangla text categorization. To summarize, the study confirms that automating the categorization of Bangla news using machine learning models is feasible. Considering the exceptional performance of this approach, it is highly recommended that it be deployed. This study establishes the foundation for further investigation and advancement in Bangla natural language processing by providing implementable remedies for news platforms and enhancing the accessibility of Bangla-language material. The findings indicate that more study and development are necessary to overcome the constraints of Bangla text categorization technology and to improve its efficiency.

CHAPTER 6

Future Work

The findings of this study have several avenues for further investigation that could improve the categorization of Bangla news headlines in the field of Bangla natural language processing (NLP). If I want my SVC model to perform better, we need to devote more time and energy to optimizing and fine-tuning it. [17] Experimenting with different kernel functions and hyperparameter settings will yield better results. Better handling of complex linguistic features in Bangla text might be possible with the help of cutting-edge algorithms and deep learning techniques, such as neural networks and transformers. To improve the classification system, addressing concerns about data imbalance is necessary. One way to do this is by using oversampling or other types of synthetic data synthesis. Possible future directions for study into incorporating contextual information into classification accuracy include using context-aware embeddings and multi-modal techniques that combine text, images, and metadata. Extending the research to include cross-domain and cross-language testing to assess the models' generalizability and adaptability is essential for a deeper understanding of their scalability. [18] Implementing the classification system into real-world applications, such as web-based and mobile app platforms, allows you to refine it and guarantee its usability. Creating web and mobile apps that leverage the categorization system can enhance the user experience by making classified news content readily available. Adding new sources and categories to the dataset can enhance the model's performance and generalizability. Through these endeavors, we will improve our ability to classify Bangla news stories and create natural language processing technologies that are effective for more minor languages.

Reference:

- [1]. Haque, A. (2018). Classification of Bangla News using Machine Learning Algorithms. *Journal of Advanced Research in Dynamical and Control Systems*, 10(6), 1516-1527.
- [2]. Hossain, M. S., & Alom, M. Z. (2018). Bangla news categorization using machine learning techniques. In 2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEICT) (pp. 1-6). IEEE.
- [3]. Saha, S., Roy, S., & Sarkar, S. (2017). Bangla news classification using machine learning techniques. In 2017 International Conference on Microelectronics, Computing & Communication Systems (MCCS) (pp. 293-296). IEEE.
- [4]. Rahman, M. M., Zaman, A. F. M. S., & Rashid, M. M. (2020). Bangla News Classification Using Machine Learning Techniques. In *Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.
- [5]. Chowdhury, P., Eumi, E.M., Sarkar, O., Ahamed, M.F. (2022). Bangla News Classification Using GloVe Vectorization, LSTM, and CNN. In: Arefin, M.S., Kaiser, M.S., Bandyopadhyay, A., Ahad, M.A.R., Ray, K. (eds) *Proceedings of the International Conference on Big Data, IoT, and Machine Learning. Lecture Notes on Data Engineering and Communications Technologies*, vol 95. Springer, Singapore. https://doi.org/10.1007/978-981-16-6636-0_54
- [6]. M. Tanvir Alam and M. Mofijul Islam, "BARD: Bangla Article Classification Using a New Comprehensive Dataset," 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, Bangladesh, 2018, pp. 1-5, doi: 10.1109/ICBSLP.2018.8554382.
- [7]. M. M. Rahman, M. A. Z. Khan and A. A. Biswas, "Bangla News Classification using Graph Convolutional Networks," 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2021, pp. 1-5, doi: 10.1109/ICCCI50826.2021.9402567.
- [8]. N. Humaira, H. Afia and S. Haque, "A Study on Machine Learning and Deep Learning Methods Using Feature Extraction for Bengali News Document Classification," 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, 2021, pp. 1-8, doi: 10.1109/ASIANCON51346.2021.9544761.
- [9]. Adrita Barua, Omar Sharif, Mohammed Moshuiul Hoque, Multi-class Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation, *Procedia Computer Science*, Volume 193, 2021, Pages 112-121, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2021.11.002>.

- [10]. Rahman, S., Chakraborty, P. (2021). Bangla Document Classification Using Deep Recurrent Neural Network with BiLSTM. In: Prateek, M., Singh, T.P., Choudhury, T., Pandey, H.M., Gia Nhu, N. (eds) Proceedings of International Conference on Machine Intelligence and Data Science Applications. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-33-4087-9_43
- [11]. 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
- [12]. 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
- [13]. H. Tan, J. Guo, and Y. Li, "E-learning Recommendation System," 2008 International Conference on Computer Science and Software Engineering, 2008, pp. 430-433
- [14]. 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]
- [15]. 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 6735-6741
- [16]. 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), Kharagpur, India, 2021, pp. 1-5, doi: 10.1109/ICCCNT51525.2021.9580116.
- [17]. N. Chy, M. H. Siddiqui and S. Das, "Bangla news classification using naive Bayes classifier," 16th Int'l Conf. Computer and Information Technology, Khulna, Bangladesh, 2014, pp. 366-371, doi: 10.1109/ICCITechn.2014.6997369. 33
- [18]. M. A. Shafin, M. M. Hasan, M. R. Alam, M. A. Mithu, A. U. Nur and M. O. Faruk, "Product Review Sentiment Analysis by Using NLP and Machine Learning in Bangla Language," 2020 23rd International Conference on Computer and Information Technology (ICCIT), DHAKA, Bangladesh, 2020, pp. 1-5, doi: 10.1109/ICCIT51783.2020.9392733.

PLAGIARISM REPORT

NewsML: A Machine Learning Approach for Bangla News Classification

ORIGINALITY REPORT

7%

SIMILARITY INDEX

9%

INTERNET SOURCES

4%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1

dspace.daffodilvarsity.edu.bd:8080

Internet Source

4%

2

link.springer.com

Internet Source

1%

3

uu.diva-portal.org

Internet Source

1%

4

B. N. Jagadesh, M. Ganesh Karthik, D. Siri, Sk. Khaja Shareef, Srihari Varma Mantena, Ramesh Vatambeti. "Segmentation using the IC2T model and Classification of Diabetic Retinopathy using the Rock Hyrax Swarm-Based Coordination Attention Mechanism", IEEE Access, 2023

Publication

1%

5

Submitted to University of Bolton

Student Paper

1%

6

d197for5662m48.cloudfront.net

Internet Source

1%

Exclude quotes On
Exclude bibliography Off

Exclude matches < 1%