Bangla News Content Classification Using Machine Learning Techniques

Towkir Ahmed	**170104110**
Md. Siam Islam	**170104124**
Popin Saha	**170104132**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Fall 2020

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2021

Bangla News Content Classification Using Machine Learning Techniques

Submitted by

Towkir Ahmed	**170104110**
Md. Siam Islam	**170104124**
Popin Saha	**170104132**

Submitted To

Faisal Muhammad Shah, Associate Professor

Farzad Ahmed, Lecturer

Md. Tanvir Rouf Shawon, Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2021

ABSTRACT

The automatic categorization of Bangla content has been explored in this research, and many supervised learning models have been presented using a huge corpus of textual data. Despite the availability of several complete textual datasets for other languages, only a few tiny datasets are prepared for the Bangla language. As a result, few research addresses Bangla document categorization, and lack of training data prevents advanced supervised learning models. There are around 77768 Bangla articles in this dataset consisting of various news from Prothom Alo. We use a variety of complex textual characteristics, such as TF-IDF, to train various supervised learning models on this immense and varied dataset. We observed that the Support Vector Machine (SVM-RBF) outperformed among all classifiers with 91.80% accuracy.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the influx of unstructured textual data, the natural language research community has become more interested in extracting insight from textual data. Text document categorization is a frequent and well-studied issue. Search, filtering, and organizing text documents are all made easier using document classification.
Several statistical and machine learning methods have been used in recent decades to extract relevant features for the correct classification of textual content. A number of features from the text data have been extracted, using TF-IDF [1] features, in order to train supervised learning models such as SVM, KNN or Naive Bayes to classify the content. Unstructured textual data has been used to extract useful features for document categorization using deep learning techniques like Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) [2] Training these supervised learning models therefore necessitates a large amount of textual information.
A few datasets are created for the categorization of Bangla documents despite the availability of numerous big datasets in other languages. As a result, only a few studies have attempted to classify Bangla text. All of these studies, however, relied on datasets of only a few hundred articles, which are insufficient for training a supervised learning algorithm. Many of today's top works, when considering how to classify Bangla content, employ simple bags of words or TF-IDF features, including semantic features-based approaches such as Word2Vec.
The Bangla Newspaper Content Dataset is a collection of Bangla documents that address the issues described above. In addition, we conduct a thorough statistical analysis to identify the textual relationship between the various types of documents. We used textual features like TF-IDF to examine the results of various supervised learning techniques, such as Linear SVM, Kernel SVM, Gaussian Naive Bayes, and Multinomial Naive Bayes.
The following summarizes the key contributions of this work:

- There are around 77768 Bangla documents in our collection, each tagged with one of

eight document categories.

- Using a variety of textual features, such as TF-IDF, we ran numerous experiments to compare the results of several supervised learning models.
- If we find any dissimilarity in score_au_presence which means dissimilarity_score of score_au_presence is not 0, then we add the apex frame number with its intensity score to dis_matrix where apex frame number and intensity score is retrieved from intensity_score's 0th and 1st index. Then we have updated intensity_score with its jth frame score_au_intensity value and updated i with jth frame as j is the dissimilar frame.

The remainder of the document is structured as follows:
In Chapter 2, we've discussed related works. Afterwards, in Chapter 3 we go through the specific features of the Bangla Newspaper dataset and the statistical text analysis of its content. As a result, the proposed Bangla content classification model is introduced in Chapter 4. In Chapter 5, we looked at how different supervised learning models performed while classifying Bangla content. Finally, in Chapter 6, we bring this project to a close by outlining possible future paths.

Chapter 2

Literature Reviews

For multi-class Bengali text categorization, a number of supervised models have been exposed.That paper [3, 4] has been thoroughly examined, particularly the data prepossessing section. They used a huge Bengali corpus with 3,76,226 articles from five distinct categories for this study and used multi-class classification to classify these five categories.For word embedding, they used Word2Vec and TF-IDF features. For our purposes, the data processing section of this study was the most important. They used the Word2Vec and TF-IDF features in the preprocessing step to accomplish text tokenization, stop word removal, and feature vector creation. In terms of the Bengali language, multi-label classification is a relatively unexplored area.

Abu Nowshed Chy et al. [5] proposed a text categorization system that uses a tree-based Naive Bayesian categorization procedure and incorporates machine learning and hierarchical structures. Because of the training feature extraction procedure and training techniques, it is a traditional machine learning system with low accuracy.

Md. Rajib Hossain et al. [6] proposed a Bangla text classification system based on machine learning, with semantic features collected from Bangla input texts using the Word2Vec method, and a documents categorization system based on multi-class SVM with SGD.For the smallest dataset, SVM and statistical machine learning-predicted algorithms produced better results than the biggest dataset.

Wahiduzzaman Akanda et al. [7] used a big dataset from one of Bangladesh's most popular Bengali newspapers, Prothom Alo, that included 4,16,289 news stories and 4,302 unique labels. Sports, Technology, Economy, Entertainment, International, and State are the six areas in which these news pieces are organized. They used Count Vectorizer for the word embedding feature. They used the ML-KNN technique and a Neural Network to create a supervised model.

Chapter 3

Data Collection & Processing

3.1 Dataset Description

The Bangla Newspaper [8] dataset contains 392772 Bangla contents, each of which is labeled with one of eight document types. This data was obtained via Kaggle. Fig. 3.1 depicts the data distribution of the gathered dataset.

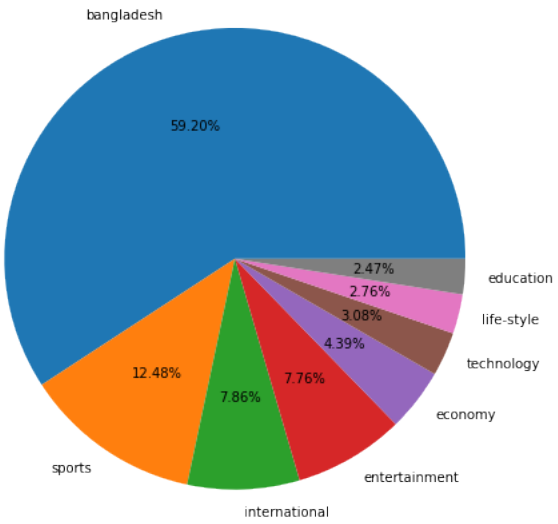


Figure 3.1: Data Distribution of the collected dataset

Since the dataset is unbalanced, we selected samples from each label that had the same length. Because the sample size of 2.47 percent in Fig. 3.1 shows it to be the lowest proportion of all, we picked it as our sample size. There are 77768 Bangla contents in total in our final dataset and the data distribution is shown in Fig. 3.2.

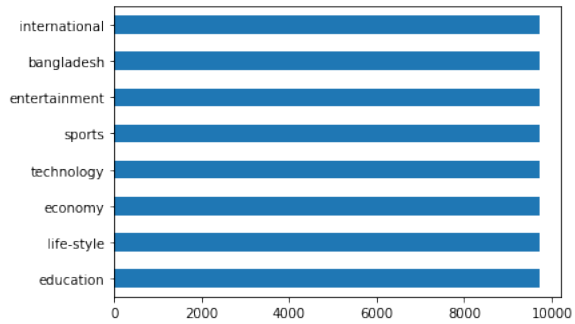


Figure 3.2: Data Distribution of the finlized dataset

3.2 Pre-Processing

During data preparation, raw data from various sources are transformed into a recognized format. The optimum training of algorithms is aided by well-preprocessed data. We have performed multi pre-processing to our given methodology.

The Bangla text documents include a large number of superfluous words that may or may not have any bearing on categorization. As a result, we must thoroughly clean the text documents. For example, removing unnecessary characters such as colon, semicolon, comma, question mark and exclamation point.

Stop words are words that have no effect on documents or sentences and are thus classified as such. Because stop words have no influence on sentences, we have eliminated them from text documents. Some of the stop words in Bengali may be found in the words are given in Fig. 3.3.

'উপর', 'উপরে', 'এ', 'এঁদের', 'এঁরা', 'এই', 'একই', 'একটি'

Figure 3.3: Stopwords in Bengali

Due to the fact that Bengali documents include a variety of punctuation, numbers, and letters, we must clean the text documents before stemming. The following Fig. 3.4 illustrates the content having various punctuation, digits and letters.

'গাজীপুরের কালিয়াকৈর উপজেলার তেলিরচালা এলাকায় আজ বৃহস্পতিবার রাতের টিফিন খেয়ে একটি পোশাক কারখানার ৫০০ শ্রমিক অসুস্থ হয়ে পড়েছেন। এ ঘটনায় বিক্ষোভ করেছেন ওই কার খানার শ্রমিকেরা।সফিপুর মডার্ন হাসপাতালের জরুরি বিভাগের চিকিৎসক আল আমিন প্রথম আলো ডটকমকে বলেন, খাদ্যে বিষক্রিয়ায় তাঁরা (শ্রমিকেরা) অসুস্থ হয়ে পড়েছেন। এতে আতঙ্কিত হওয়ার কিছু নেই। অসুস্থদের চিকিৎসা দেওয়া হয়েছে।কারখানার শ্রমিক ও পুলিশ সূত্রে জানা যায়, উপজেলার তেলিরচালা এলাকার সেজাদ সোয়েটার লিমিটেড কারখানার শ্রমিকদের আজ রাত সাড়ে সাতটার দিকে টিফিন দেওয়া হয়। টিফিনে ছিল ডিম, রুটি, পেটিস ও কলা। টিফিন খেয়ে শ্রমিকেরা যথারীতি কাজে যোগ দেন। ওই টিফিন খাওয়ার প্রায় এক ঘণ্টা পর রাত সাড়ে আটটার দিকে কয়েকজন শ্র মিকের বমি ও পেট ব্যথা শুরু হয়। এরপর ধীরে ধীরে পুরো কারখানার শ্রমিকেরা অসুস্থ হতে থাকে। অনেকেই কারখানার মেঝেতে ঢলে পড়ে। এতে পাঁচ শতাধিক শ্রমিক অসুস্থ হয়ে পড়ে।পরে কার খানা কর্তৃপক্ষ দ্রুত যানবাহনের ব্যবস্থা করে তাদের সফিপুর জেনারেল হাসপাতাল, সফিপুর মডার্ন হাসপাতাল, উপজেলা স্বাস্থ্য কমপ্লেক্সসহ বিভিন্ন ক্লিনিকে ভর্তি করে।...'

Figure 3.4: The content of Bengali documents before cleaning

We tokenized the text documents after cleaning them by utilizing the space character as a delimiter. Because our text documents are not normalized, we must normalize them. Within the area of Natural Language Processing, there are two distinct methods for normalizing text that are used to prepare text, words, and documents for further processing. We utilized stemming to reduce a word to its root, also known as a lemma, which affixes to suffixes and prefixes or to word roots. We used the BNLP toolkit [9] to stem the tokenized words. Fig. 3.5 depicts the stemmed content of the text documents. Fig. 5.5 illustrates the data statistics for the preprocessed texts.

'গাজীপুর কালিয়াকৈর উপজেল তেলিরচালা এলাকায় বৃহস্পতিব রাত টিফিন খেয়ে পোশাক কারখ শ্রমিক অসুস্থ পড়েছেন ঘটনায় বিক্ষোভ কারখ শ্রমিকেরা সফিপুর মডার্ন হাসপাতাল জরুরি বিভাগ চিকিত্সক আল আমিন আলো ডটকম খাদ্যে বিষক্রিয়ায় শ্রমিকেরা অসুস্থ পড়েছেন আতঙ্কিত অসুস্থ চিকিত্সা কারখ শ্রমিক পুলিশ সূত্রে উপজেল তেলিরচালা এল সেজাদ সোয়ে লিমিটেড কারখ শ্রমিক রাত সাড়ে সাত টিফিন টিফিনে ডিম রুটি পেট কলা টিফিন খেয়ে শ্রমিকেরা যথারীতি যোগ টিফিন খাওয় এক ঘণ্ রাত সাড়ে আট কয়েকজন শ্রমিক বমি পেট ব্যথা এরপর ধীরে ধীরে পুরো কারখ শ্র মিকেরা অসুস্থ কারখ মাঝ ঢলে পড়ে পাঁচ শতাধিক শ্রমিক অসুস্থ পড়ে কারখানা কর্তৃপক্ষ দ্রুত যানবাহন ব্যবস্থা সফিপুর জেনারেল হাসপাতাল সফিপুর মডার্ন হাসপাতাল উপজেলা স্বাস্থ্য কমপ্লেক্সসহ ক্লিনি ভর্তি বাসি পচা খাবা দেওয়ায় শ্রমিক ক্ষুব্ধ কারখ বিক্ষোভ খবর পুলিশ শ্রমিক বুঝিয়ে খাবা সরবরাহ প্রতিষ্ঠান বিরুদ্ধে ব্যবস্থা আশ্বাস দিলে শ্রমিকেরা শান্ত সফিপুর জেনারেল হাসপাতালে ভর্তি শ্র মিক জাকির হোস আসমা আক্ত টিফিন খাওয় সময় ডিম কেক দুর্গন্ধ বের হচ্ছিল কারণে খাবা খায়নি বেশির ভাগ শ্রমিক খাবা খেয় কারখ সহকা...'

Figure 3.5: The content of Bengali documents after stemming

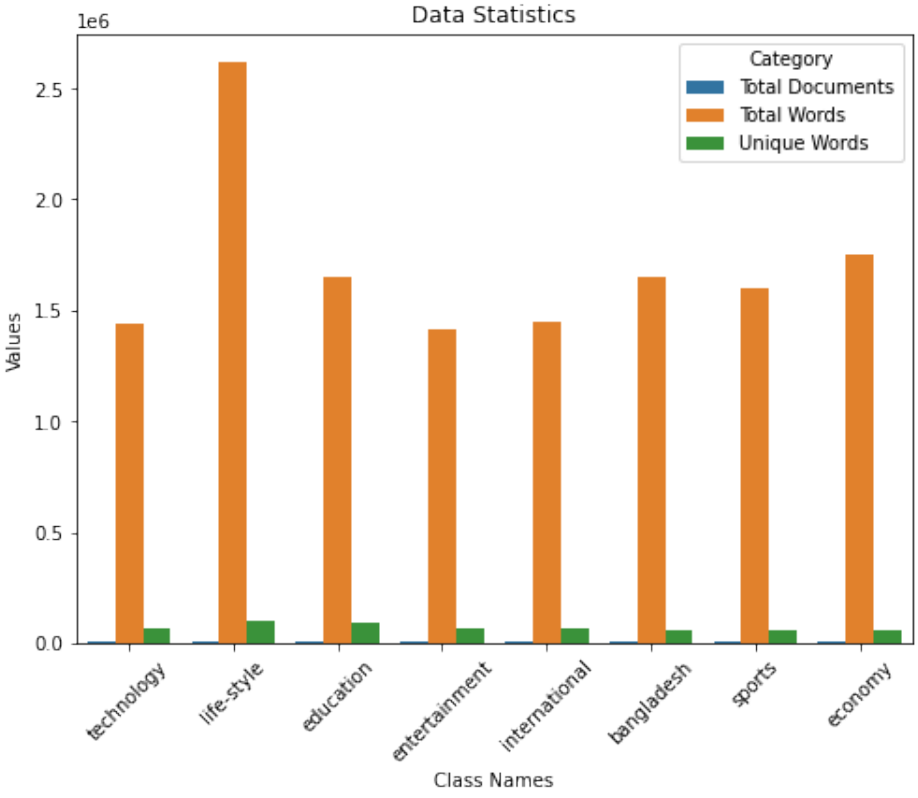


Figure 3.6: Statistical distribution of pre-processed data

3.3 Encoding

Due to the fact that our dataset contains categorical labels, we must convert it to a numerical format for machine learning implementation. There are two methods for categorical encoding: label encoding and one-hot encoding. We utilized label-encoding to encode our category labels in our study. This is a fairly straightforward method that entails turning each value in a column to a number. The encoded labels are shown in Fig. 3.7.

	Label before encoding	Label after encoding
0	bangladesh	0
1	economy	1
2	education	2
3	entertainment	3
4	international	4
5	life-style	5
6	sports	6
7	technology	7

Figure 3.7: Label encoding

Chapter 4

Methodology

The proposed framework relies on text modality. The details of the proposed method were depicted in Fig. 4.1. The whole workflow is divided into four phases to classify Bangla news content, including data acquisition, data processing, and data classification using the proposed machine learning classifiers.

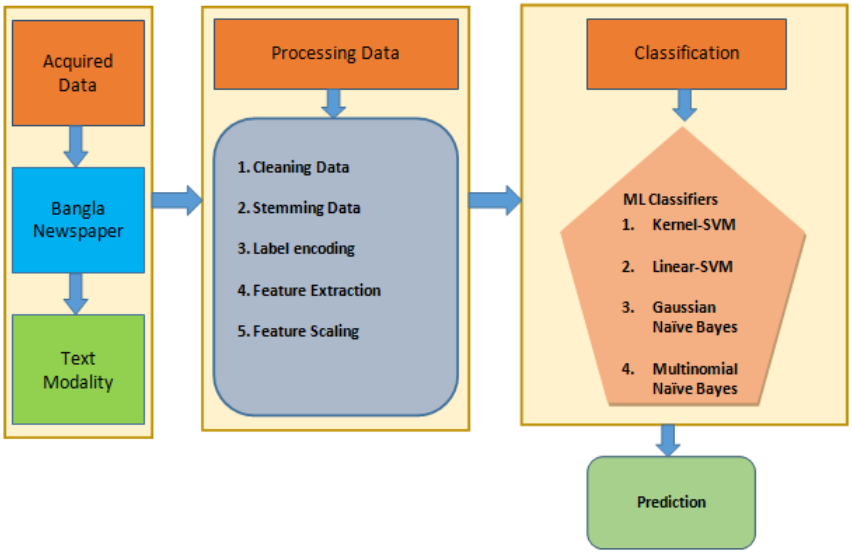
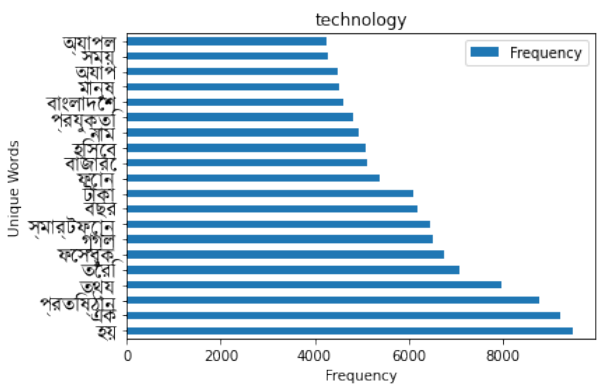


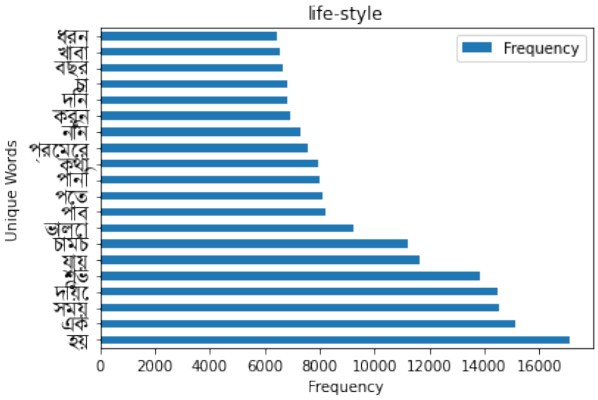
Figure 4.1: Diagram of proposed methodology

4.1 Feature Extraction

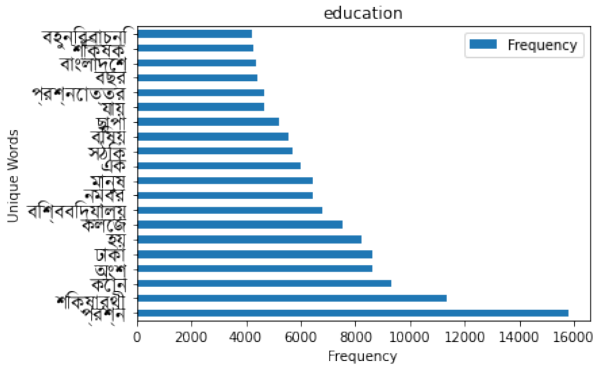
After finishing the pre-processing procedures, we used the TF-IDF method to extract the feature vector.
After that, we trained various supervised classifiers, which are detailed in the following sections.



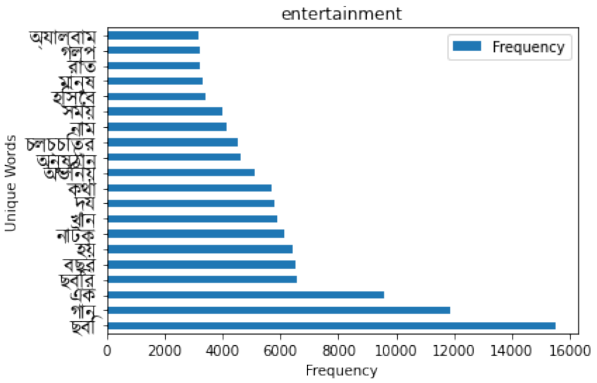
(a) Most frequent words of technology category



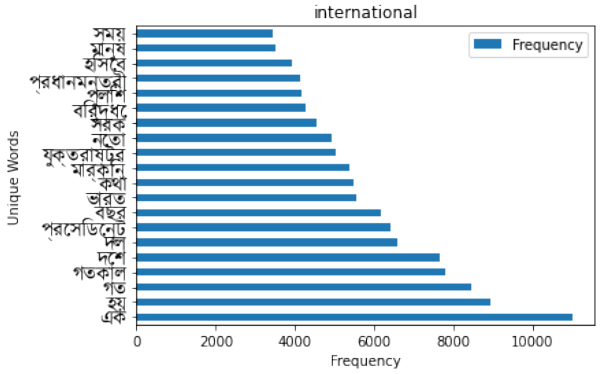
(b) Most frequent words of life-style category



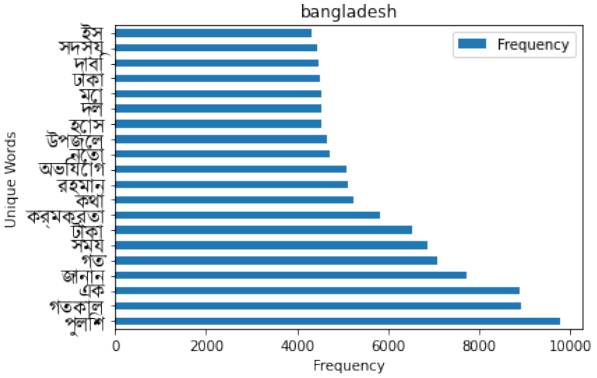
(c) Most frequent words of education category



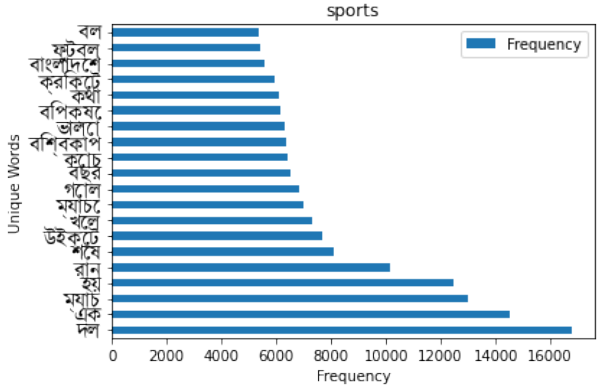
(d) Most frequent words of entertainment category



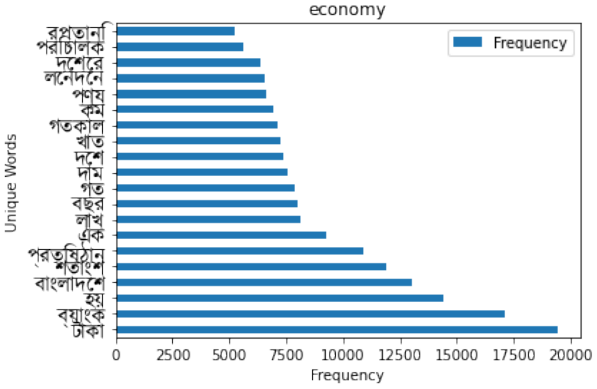
(e) Most frequent words of internation category



(f) Most frequent words of bangladesh category



(g) Most frequent words of sports category



(h) Most frequent words of economy category

Figure 4.2: Most frequent words after pre-precossing data

1. TF-IDF Feature: TF-IDF [13] is a method that is often used in Information Retrieval. It is a method of giving numerical values to words in a text. The TF-IDF score for a word w in a corpus document D may be computed as follows:
   $TF = \frac{Frequency\ of\ w\ in\ D}{Total\ number\ of\ words\ D}$, $IDF = \log_e \frac{Total\ number\ of\ documents}{Number\ of\ documents\ containing\ w}$
   $TF - IDF = TF * IDF$
   We extracted a total of 3500 sized feature vector. After extracting these features, we used them to train multiple supervised learning models. Then, we selected the best model for the TF-IDF features and presented a comparison of all models.

4.2 Feature Scaling

Most of the time, the dataset does not stay on the same scale, and it is not even standardized. As a result, feature scaling is a basic data transformation technique for adapting a dataset to algorithms. To achieve the same scale for all data, we must scale the value of features and provide equal weight to all features. Furthermore, scaling may vary in various values for different features. There are many feature scaling methods, such as standardization, mean normalization, min-max scaling, unit vector, and so on.
As the features are limited inside a defined region, we used the Standardization technique in our study. Standardization ensures that values are centered on the mean and have a single standard deviation. This results in a zero mean for the attribute and a unit standard deviation for the resulting distribution.

4.3 Machine Learning Classifiers

We implemented four machine learning classifiers, including Linear-SVM, Kernel-SVM, Gaussian Naive Bayes, and Multinomial Naive Bayes, since we are primarily concerned with the performance of Bangla News Classification. The purpose of this experiment is to determine which classifier is the best for classifying Bangla news content for the test-sets. We have used 70% of data as training-set and 30% of data as test-set. The following subsections detail the whole process for the experiment.

4.3.1 Linear SVM

The most applicable machine learning algorithm for our problem is Linear SVC. The objective of a Linear SVC (Support Vector Classifier) is to fit the data we provide, returning a "best fit" hyperplane that divides, or categorizes, our data. From there, after getting the

hyperplane, we can then feed some features to the classifier to see what the 'predicted' class is. This makes this specific algorithm rather suitable for classification.

4.3.2 Non-linear SVM

Non-Linear SVM may also be used to solve our problem. SVM supports a variety of non-linear kernel types, including RBF, Polynomial, Sigmoid, and so on. We classified our data using the Radial Basis Function (RBF) kernel. Non-linear kernels are advantageous for data that cannot be separated by a linear hyperplane. This kernel is used to convert non-separable data to separable data. This kernel classifies data by mapping it onto a high-dimensional space.

4.3.3 Gaussian Naive Bayes

Gaussian Naive Bayes algorithm is also suitable for text classification. Gaussian Naive Bayes is a Naive Bayes version that uses the Gaussian normal distribution and works with continuous data. To build a basic model quickly, suppose the data is represented by a Gaussian distribution with no co-variance across dimensions. The mean and standard deviation of the points inside each label are used to fit this model.

4.3.4 Multinomial Naive Bayes

The Multinomial Naive Bayes algorithm is a probabilistic learning technique used to categorize text documents in Natural Language Processing (NLP). The program guesses the tag of a text, such as newspaper content, using the Bayes theorem. It computes the likelihood of each tag for a given sample and outputs the tag with the greatest likelihood.

Chapter 5

Experiments and Results

5.1 Experimental Tool

The entire operation was carried out in the Anaconda distribution using the Python 3.7.4 programming language. The Python library includes several tools for implementing machine learning. Pandas is an unrivaled data representation library with vast commands and data management capabilities. We've used it to read and analyze data. Scikit-learn allows us to create models using a variety of classification algorithms. By training and testing machine learning models using the best feasible feature sets as input and labels as output, we utilize matplotlib to visualize their performance. We have used the BNLP toolkit [9] to pre-process Bangla text documents.

5.2 Results

After applying four machine learning algorithms, we got the following results:

	precision	recall	f1-score	support
0	0.85	0.90	0.87	2918
1	0.85	0.86	0.85	2948
2	0.90	0.92	0.91	2882
3	0.90	0.90	0.90	2998
4	0.89	0.88	0.89	2911
5	0.85	0.86	0.85	2914
6	0.95	0.93	0.94	2829
7	0.88	0.81	0.84	2931
accuracy			0.88	23331
macro avg	0.88	0.88	0.88	23331
weighted avg	0.88	0.88	0.88	23331

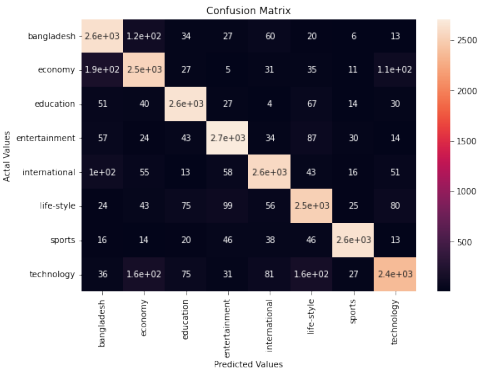


Figure 5.1: Classification report and confusion matrix for Linear SVM

	precision	recall	f1-score	support
0	0.92	0.90	0.91	2918
1	0.88	0.91	0.90	2948
2	0.94	0.93	0.93	2882
3	0.93	0.94	0.93	2998
4	0.90	0.94	0.92	2911
5	0.91	0.89	0.90	2914
6	0.96	0.96	0.96	2829
7	0.90	0.89	0.89	2931
accuracy			0.92	23331
macro avg	0.92	0.92	0.92	23331
weighted avg	0.92	0.92	0.92	23331

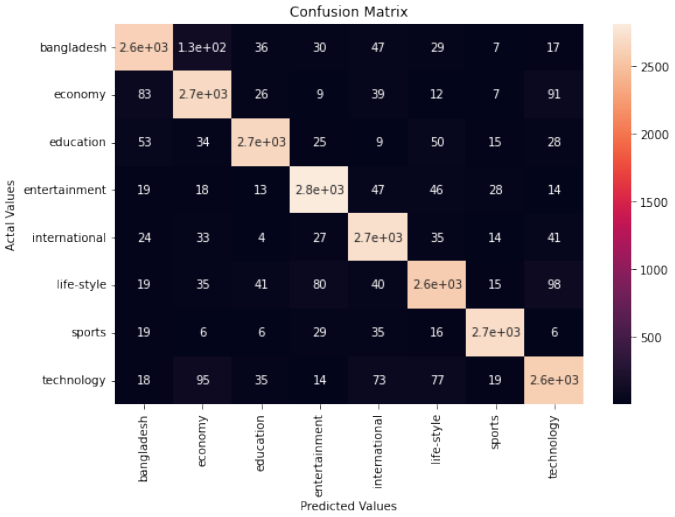


Figure 5.2: Classification report and confusion matrix for RBF-SVM

	precision	recall	f1-score	support
0	0.67	0.94	0.78	2918
1	0.85	0.70	0.77	2948
2	0.93	0.85	0.89	2882
3	0.79	0.82	0.81	2998
4	0.78	0.83	0.80	2911
5	0.91	0.61	0.73	2914
6	0.83	0.89	0.86	2829
7	0.77	0.79	0.78	2931
accuracy			0.80	23331
macro avg	0.82	0.80	0.80	23331
weighted avg	0.82	0.80	0.80	23331

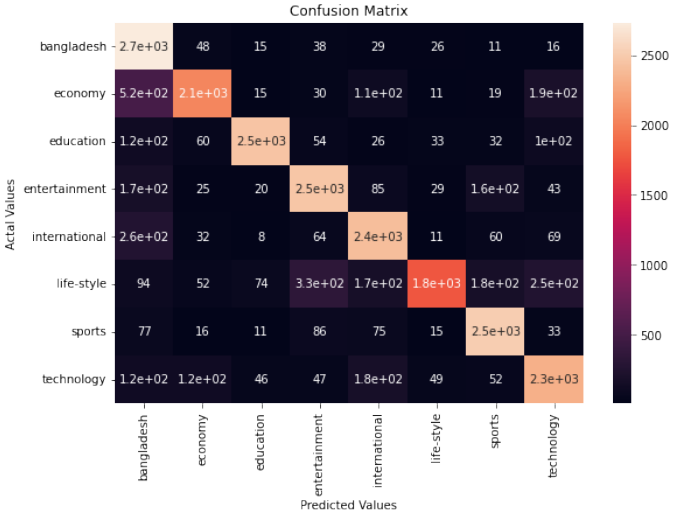


Figure 5.3: Classification report and confusion matrix for Gaussian Naive Bayes

Table 5.1: Comparison of the outcomes achieved using various classifiers

Method	*Accuracy*	*Precision*	*Recall*	*f1-Score*
RBF-SVM	91.80	91.84	91.81	91.81
Gaussian-NB	80.21	81.62	80.26	80.13
Multinomial-NB	88.12	88.34	88.13	88.16
Linear-SVM	88.38	88.47	88.40	88.40

	precision	recall	f1-score	support
0	0.88	0.89	0.88	2918
1	0.84	0.86	0.85	2948
2	0.94	0.86	0.90	2882
3	0.87	0.91	0.89	2998
4	0.88	0.91	0.89	2911
5	0.89	0.81	0.85	2914
6	0.95	0.92	0.94	2829
7	0.82	0.88	0.85	2931
accuracy			0.88	23331
macro avg	0.88	0.88	0.88	23331
weighted avg	0.88	0.88	0.88	23331

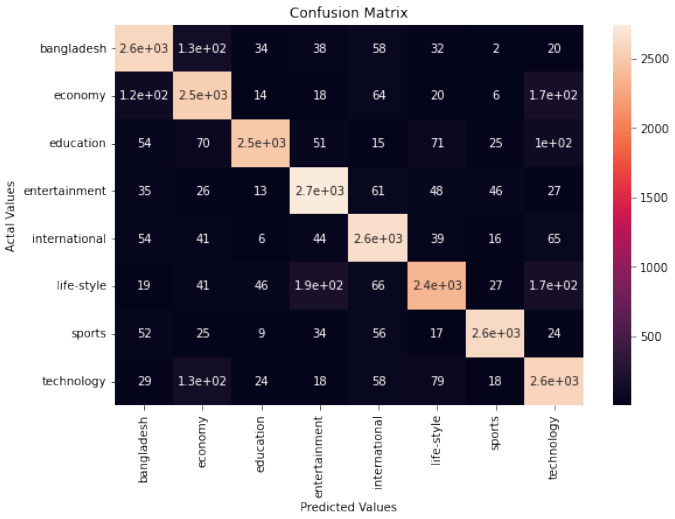


Figure 5.4: Classification report and confusion matrix for Multinomial Naive Bayes

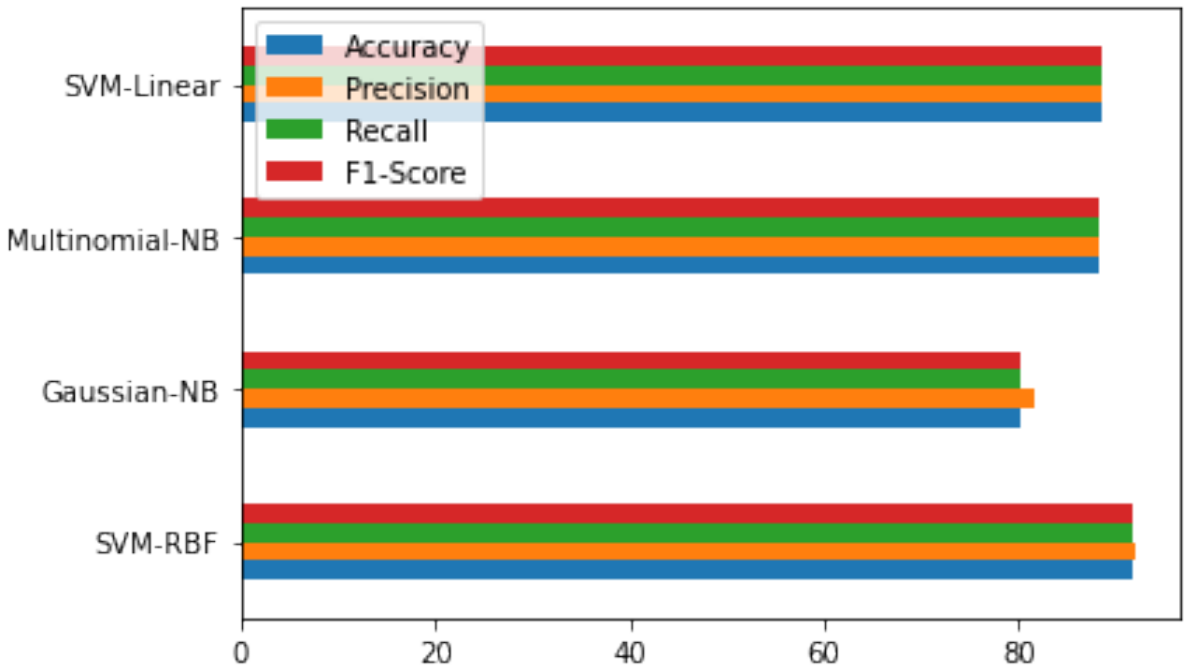


Figure 5.5: Bar plot for performance metrics of various machine learning classifiers

Chapter 6

Future Work and Conclusion

The growing domain of online newspapers presents a rich area, which can benefit immensely from the automatic classification approach. In this project we present a system of automatically classifying Bangla News documents. This system has the capability to provide users with efficient and reliable access to classified news from different sources. We used Linear-SVM, RBF-SVM, Gaussian Naïve-Bayes, Multinomial Naïve-Bayes. Among them, RBF SVM performed best with an accuracy of 91.80%.

As we have taken a small sample size, we have a plan to extend this dataset in the future so that it can be used to solve other Bangla NLP related problems. Furthermore, deep learning models such as LSTM, BiLSTM, and BERT can also be considered to improve the prediction model.

References

[1] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[3] A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, "Performance of classifiers in bangla text categorization," in *2018 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pp. 168–173, IEEE, 2018.

[4] M. T. Alam and M. M. Islam, "Bard: Bangla article classification using a new comprehensive dataset," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–5, IEEE, 2018.

[5] A. N. Chy, M. H. Seddiqui, and S. Das, "Bangla news classification using naive bayes classifier," in *16th Int'l Conf. Computer and Information Technology*, pp. 366–371, IEEE, 2014.

[6] M. R. Hossain and M. M. Hoque, "Automatic bengali document categorization based on word embedding and statistical learning approaches," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pp. 1–6, IEEE, 2018.

[7] W. Akanda and A. Uddin, "Multi-label bengali article classification using ml-knn algorithm and neural network," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, pp. 466–471, IEEE, 2021.

[8] Z. A. N. Nabil, "Bangla Newspaper Dataset." https://www.kaggle.com/furcifer/bangla-newspaper-dataset. [Accessed: 21 October, 2020].

[9] S. Sarker, "Bengali Natural Language Processing(BNLP)." https://github.com/sagorbrur/bnlp.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Tuesday 28th September, 2021 at 7:03pm.