Multi-category Bangla News Classification using Machine Learning Classifiers and Multi-layer Dense Neural Network

Sharmin Yeasmin[1], Ratnadip Kuri[2], A R M Mahamudul Hasan Rana[3*]
Ashraf Uddin[4], A. Q. M. Sala Uddin Pathan[5], Hasnat Riaz[6]
Department of Computer Science and Telecommunication Engineering
Noakhali Science and Technology University, Bangladesh[1, 2, 3, 5, 6]
Department of Computer Science, American International University-Bangladesh, Bangladesh[4]

***Abstract*—Online and offline newspaper articles have become an integral phenomenon to our society. News articles have a significant impact on our personal and social activities but picking a piece of an appropriate news article is a challenging task for users from the ocean of sources. Recommending the appropriate news category helps find desired articles for the readers but categorizing news article manually is laborious, sluggish and expensive. Moreover, it gets more difficult when considering a resource-insufficient language like Bengali which is the fourth most spoken language of the world. However, very few approaches have been proposed for categorizing Bangla news articles where few machine learning algorithms were applied with limited resources. In this paper, we accentuate multiple machine learning approaches including a neural network to categorize Bangla news articles for two different datasets. News articles have been collected from the popular Bengali newspaper Prothom Alo to build Dataset I and dataset II has been gathered from the famous machine learning competition platform Kaggle. We develop a modified stop-word set and apply it in the preprocessing stage which leads to significant improvement in the performance. Our result shows that the Multi-layer Neural network, Naïve Bayes and support vector machine provide better performance. Accuracy of 94.99%, 94.60%, 95.50% has been achieved for SVM, Logistic regression and Multi-layer dense neural network, respectively.**

Keywords*—Bangla news classification; supervised learning; feature extraction; category prediction; machine learning; neural network

I. Introduction

A newspaper is known as a powerhouse of information. People get the latest information about their desired content through online or offline newspapers. Thousands of newspapers are published in different languages all over the world. Whatever happens around the world may be a thousand miles away but reaches us within a second through online news content. In the recent years, the importance of online articles has also increased rapidly due to the rapid rise and availability of smart devices. Bangla is the fourth most spoken language and vast amounts of Bangla news articles are produced every hour worldwide. Choosing the appropriate information from the sea of web is difficult as the news has no categorization based on its content. Online news websites provide subject categories and sub-categories [1] which significantly vary newspaper to newspaper. So, these might not be sufficient for fulfilling users' choice of interest. Readers like to explore news from various news sources rather than one source and recommending suitable news to the readers based on its contents can improve the readers' experience.

The paper's main motivation is to help in recommending relevant news to the Bengali online news readers using multi-category classification. Readers are only attracted to the news articles of their interest [2]. For this purpose, the readers have to explore all the news articles of different news sites to get the desired items. For example, a user interested in entertainment-related news has to go through all the news articles from various news sites and analyze information from multiple tiresome sources. A user would prefer such a system or framework that would gather news articles of interest from various news sites and access the system anywhere on any electronic device. Although frameworks are available to notify the readers about news' on their desire categories, manually categorizing thousands of online Bangla news articles is challenging. Moreover, appropriate categorization of Bangla news articles considering their content is essential for the readers and designing an automated system for this purpose is a crying need.

Several approaches have been proposed for news categorization for different languages, i.e. Indonesian [4], Hindi[5], Arabic[6][11], Spanish [7], and these approaches mainly based on traditional machine learning algorithms such as Naïve Bayes, decision tree, K-Nearest Neighbors etc. Since Bengali is morphologically rich and complex considering the large scale of alphabets, grapheme and dialects, it needs special consideration of its features in the training phase for classification on Bangla news based on its context. However, some approaches are available in Bangla language [13-16], but these researches were limited to some traditional methods and dealt with small datasets. Due to the scarcity of resources and the complex structure of Bangla text, it's been a challenging task to classify the Bangla news.

In this paper several popular machine learning models and a multi-layer dense neural network are implemented on two different datasets. Dataset I has been built of five categories called Economics, Entertainment, International, Science and Technology and, Sports containing 1425 documents from

*Corresponding Author

popular Bangla newspaper Prothom Alo available on [20] and collected a dataset named dataset II from the Kaggle website [17] which has a total of 532509 records with nineteen categories. But, 169791 records of five categories are used from that dataset in this paper. A list of Bangla stop words are built containing 875 words [21] to remove from the newspaper contents for preprocessing purpose. Similar preprocessing steps are applied for both datasets separately and achieved better accuracy for multiple machine learning models. The accuracy of 92.63% and 95.50% for dataset I and dataset II was achieved for the multi-layer dense neural network, respectively.

The remaining part of the paper is organized as follows - Section II reviews several related works on different types of news classification both for Bangla and other languages. Section III presents research methodology which describes datasets and proposed methods. Section IV depicts result analysis. Finally, this work is concluded and provides future direction in Section V.

II. Related Work

Text classification is the process of assigning labels to text according to its content. It is one of the most fundamental tasks in Natural Language Processing (NLP) with broad application such as sentiment analysis, topic labeling, spam detection, intent detection etc. Nowadays, many tasks have been conducted on this field. Especially it is done for English language as there are enough resources for English language [3]. On the other hand, there are not enough resources except English for the task because very few works have been carried out for the task. However, working on this field is also increasing day by day in recent times. Some works of text classification on non-English languages are overviewed in the following:

Naïve Bayes and Two-Phase Feature Selection Model were used to predict the test sample category for Indonesian news classification. Naive Bayes classifier is quicker and efficient than the other discriminative models. In text classification applications and experiments, Naive Bayes (Naïve Bayes) probabilistic classifier is often used because of its simplicity and effectiveness using the joint probabilities of words and categories given a document [4]. M. Ali Fauzi et al. [4] used Naïve Bayes for Indonesian news classification. Abu Nowshed Chy et al. [10] used Naïve Bayes for Bangla news classification.

Machine learning approach was used for the classification of indirect anaphora in Hindi corpus [5]. The direct anaphora has the ability to find the noun phrase antecedent within a sentence or across few sentences. But, indirect anaphora does not have explicit referent in the discourse. They suggested looking for certain patterns following the indirect anaphora and marking demonstrative pronoun as directly or indirectly anaphoric accordingly. Their focus of study was pronouns without noun phrase antecedent.

A method was designed for classification of Arabic news, the classification system that best fits data given a certain representation [6]. A new method was presented for Arabic news classification using field association words (FA words). The document preprocessing system generated the meaningful terms based on Arabic corpus and Arabic language dictionary. Then, the field association terms were classified according to FA word classification algorithm. It is customary for people to identify the field of document when they notice peculiar words. These peculiar words are referred to as Field Associating words (FA words); specifically, they are words that allow us to recognize intuitively a field of text or field-coherent passage. Therefore, to identify the field of a passage FA terms can be used, and to classify various fields among passages FA terms can be also used.

Cervino U et al. applied machine learning techniques to the automatic classification of news articles from the local newspaper La Capitaolf Rosario, Argentina [7]. The corpus (LCC) is an archive of approximately 75,000 manually categorized articles in Spanish published in 1991. They benchmarked on LCC using three widely used supervised learning methods: k-Nearest Neighbors, Naive Bayes and Artificial Neural Networks, illustrating the corpus properties.

This paper delineates the Bangla Document Categorization using Stochastic Gradient Descent (SGD) classifier [8]. Here, document categorization is the task in which text documents are classified into one or more of predefined classes based on their contents using Support Vector Machines and Logistic Regression. Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning. In text classification and natural language processing, SGD has been successfully applied to large-scale and sparse machine learning problems often encountered.

Fouzi Harrag, Eyas EI Qawasmah [11] used ANN for the classification of Arabic language document. In this paper Singular Value Decomposition (SVD) had been used to select the most relevant features for the classification.

Neural network was used for web page classification based on augmented PCA [12]. In this paper, each news web page was represented by term weighting schema. The principal component analysis (PCA) had been used to select the most relevant features for the classification. Then, the final output of the PCA is augmented with the feature vectors from the class-profile which contains the most regular words in each class before feeding them to the neural networks. According to this paper it's evident that, in case of Sports news, WPCM provides most acceptable classification accuracy based on their datasets. Their experiment evaluation also demonstrates the same.

A research group of Shahjalal University of Science & Technology used different machine learning based approaches of baseline and deep learning models for Bengali news categorization [13]. They used baseline models such as: Naïve Bayes, Logistic Regression, Random Forest and Linear SVM and deep learning models like BiLSTM, CNN. They found out that the highest result comes from the Support Vector Machine in the base model and CNN in deep learning where CNN gave the best performance for their Dataset.

In paper [14] authors used multi-layer dense neural network for Bangla document categorization. As feature selection technique they used TF-IDF method. They used three dense layers and 2 dropout layers. They got 85.208% accuracy.

Authors on [15] used four supervised learning methods namely Decision Tree, K-Nearest Neighbor, Naïve Bayes, and Support Vector Machine for categorization of Bangla web documents. They also build their own dataset corpus but they didn't publish it. Their corpus included 1000 documents with a total number of words being 22,218. Their Dataset included five categories such as business, health, technology, sports and education. As feature selection they used TF-IDF method and they got 85.22% f-measure for Naïve Bayes, 74.24% for K-Nearest Neighbor, 80.65% for Decision Tree and 89.14% for Support Vector Machine.

An exploration group used Bidirectional Long Short Term Memory (BiLSTM) for classification of Bangla news articles [16]. They used Gensim and fastText model for vectorization of their text. Their Dataset contained around 1 million articles and 8 different categories. They got 85.14% accuracy for BiLSTM for their Dataset.

III. Methodology

The goal of this proposed model is to categorize Bangla news automatically based on the content of the document. In order to meet this up, some steps are performed such as 1) Data collection, 2) Data preprocessing, 3) Feature selection and extraction, 4) Dividing Dataset into training and testing set, 5) Building and fitting models, 6) Category prediction. Fig. 1 depicts an overview of the approach. The details of the steps are explained in following paragraphs.

A. Data Collection

Data is crucial in machine learning which required a lot of data to come up with somewhat generalizable models. The Bangla dataset corpus is built for this research task & the news articles have been collected from the popular news portal Prothom Alo online newspaper. News articles of five categories such as 'International', 'Economics', 'Entertainment', 'Sports', 'Science and Technology' has been used for the dataset. This dataset corpus consists of 1425 documents. Each category contains 285 documents, which can be found at [20]. Details of the dataset are represented in Table I.

Another dataset is also downloaded from the Kaggle website [17]. This Dataset contains newspaper articles from 2013 to 2019 from Prothom Alo. The newspaper articles have already been classified into different categories such as International, State, Economy, etc. Only five categories, namely, Entertainment, International, Economic, Sports, and Technology. In the Table II, the details and statistical analysis of the whole Dataset are given.

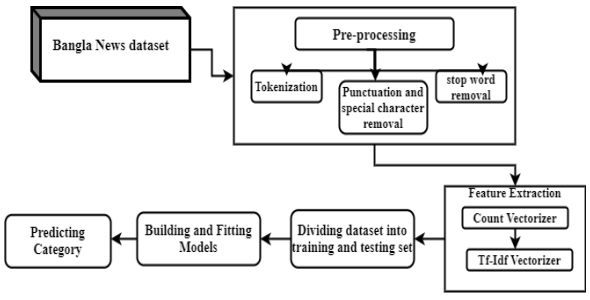


Fig. 1. Overview of Bangla News Classification System.

TABLE I. Details of Dataset I

Category	No. of Docs	Words/Doc(Average)
Economics	285	433
Entertainment	285	380
International	285	299
Science &Technology	285	381
Sports	285	349
Total	1425	367

TABLE II. Details of Dataset II

Category	No. of Docs	Words/Doc
Economics	20858	277
Entertainment	36791	237
International	37176	235
Science & Tech	15117	231
Sports	59849	261
Total	169791	250

B. Data Pre-processing

Data Preprocessing is a technique that is used to convert the raw data into a clean data set. Preprocessing the data is an important task and it is essential for getting better accuracy. In the experiment, the data was processed by several techniques such as removing empty data from document, tokenization, punctuation removal and stop word removal, white space removal, number removal.

1) Tokenization: Splitting a text into sentences, then words, and then characters. Based on spaces, texts are broken down into words and using the list function; words are broken down into characters.

2) Punctuation, special character and number removal: Punctuation like ; : । ‘ " ’ , ? !, etc. and special character like @, #, $, %, ^, &, (, *,), etc. and number that is not important for classification are removed from the whole Dataset.

3) Stop word removal: High-frequency words common in every document and have not much influence in the text are called stop words. Stop words are collected from two different sources [18]&[19], and combined unique stop words and increased the number of stopwords. The stop words list that was build contains 875 stop words, and it can be found at [21]. The list of 361 bangali stop words like "অবশ্য, অনেকে, এ, এবং, ইত্যাদি, করেছিলেন, নিতে, হয়, etc." All the stop words are removed from the Dataset for getting better accuracy.

4) Categorical encoding: There are two types of categorical encoding entitled label encoding and one-hot encoding. In label encoding, each label is assigned a unique integer based on alphabetical ordering. On the other hand, each category is represented as a one-hot vector in one hot encoding. That means only one bit is hot or true at a time. An example of a one-hot encoding of a dataset with two categories is given in Table III. Label encoding technique has been used for encoding category in machine learning algorithms and one-hot encoding for multi-layer dense neural network.

TABLE III. EXAMPLE OF ONE HOT ENCODING

	Label 1	Label 2	Label 3
Doc 1	0	0	1
Doc 2	1	0	0

After the data preprocessing step, statistical analysis step is performed on both Dataset to see if data preprocessing step is successfully performed and how words are related to each category. Fig. 2 illustrates the flow chart of the data preprocessing system. Fig. 3 and Fig. 4 illustrate the 14 most frequent words of each category of Dataset I and Dataset II. It is seen that these words are strongly related to corresponding categories that help the model successfully predict a document category. After pre-processing step, structure and number of word is changed on datasets. The detailing after pre-processing step of the two dataset is given in Table IV and Table V.

C. Feature Selection and Extraction

In this step, string features are converted into numerical features. Bag of words and TF-IDF model are used for converting string features into numerical features for performing the mathematical operation. Dataset I consists of 43404 unique words, and Dataset II that is downloaded from the Kaggle website [17] consists of 915428 unique words after data preprocessing. All the words do not have impact on the classification. So, the most frequent words have been used as features that have importance to classification. For selecting features, a Count vectorizer was utilized, which works based on the frequencies of words. Both datasets' model accuracy are observed in the Count vectorizer approach by considering different minimum document frequencies and maximum document frequencies. And for Dataset I, the best result is found by considering minimum document frequency 10, which means the words are excluded that are only on 10 or less than 10 documents and maximum document frequency 0.6, which means the words that are on the 60% document or more than that. For Dataset II, the highest accuracy is got by considering minimum frequency 10 and maximum document frequency 0.6 because those words have no significance in determining the class. In this paper 1320 most frequent words are used as a feature vector for Dataset I, and the rest of the words are excluded. For Dataset II, 10,000 most frequent words are used as a feature vector.

After selecting features, the TF-IDF vectorizer has been used for feature extraction because count vectorizer doesn't return the proper value. As it is known, count vectorizer only returns 0 or 1 as the value of a different word which does not states the transparent frequency of different words from a document.

1) Bag of words: It is a basic model used in natural language processing. A bag-of-words is a representation of text that describes the occurrence of words within a document.

2) TF-IDF: TF-IDF stands for Term Frequency-Inverse Document Frequency which says the word's importance in the corpus or Dataset. TF-IDF contain two concept Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency is defined as how frequently the word appears in the document or corpus. Term frequency can be defined as:

TF = No. of time word appear in the doc. / Total no. of word in the doc.

Inverse document frequency is another concept that is used for finding out the importance of the word. It is based on the fact that less frequent words are more informative and essential. IDF is represented by the formula:

IDF = No of Docs / No of Docs in which the word appears

TF-IDF is a multiplication between TF and IDF value. It reduces the importance of the common word that is used in a different document. And only take important words that are used in classification. TF-IDF matrix of first 10 docs and first six words of dataset I is given in Table VI.

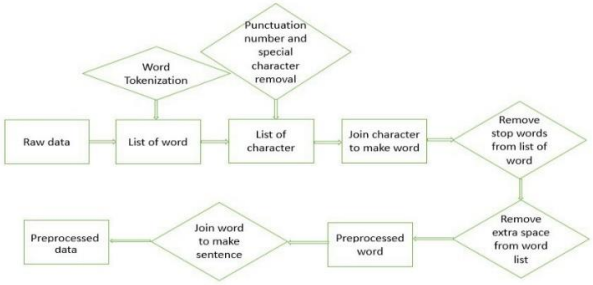


Fig. 2. Flow Chart of Data Preprocessing System.

TABLE IV. DATASET I DETAILS AFTER PREPROCESSING

Category	Total Words	Words/Doc	Unique Words
Economics	75379	282	12462
Entertainment	64895	243	15722
International	53174	199	11883
Science and Technology	65830	244	14381
Sports	62449	230	13300
Total	341702	240	55748

TABLE V. DATASET II DETAILS AFTER PREPROCESSING

Category	**Total Words**	**Words/Doc**	**Unique Words**
Economics	5071921	243	271773
Entertainment	4972334	135	290961
International	3319266	89	158302
Science and Technology	1971826	130	147524
Sports	8879569	148	379027
Total	24214916	142	1247587

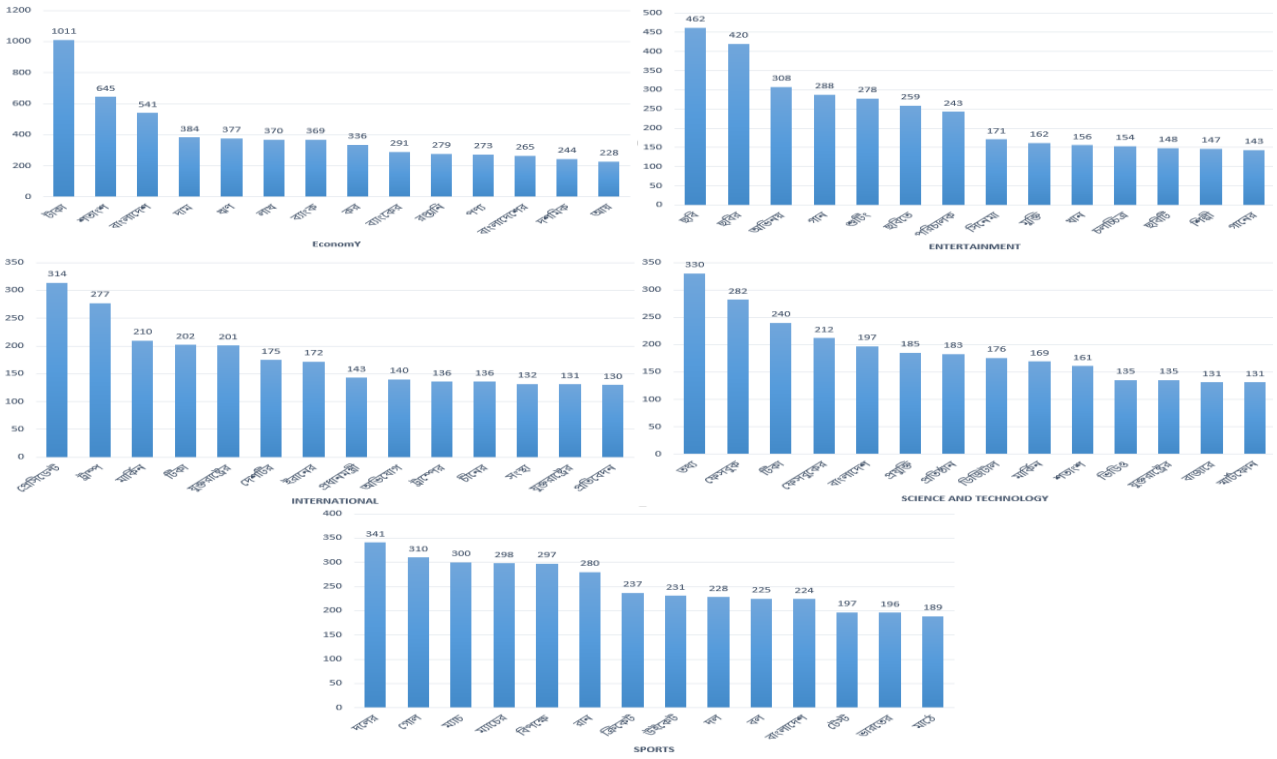


Fig. 3. Fourteen most frequent words of each category after data cleaning of Dataset I

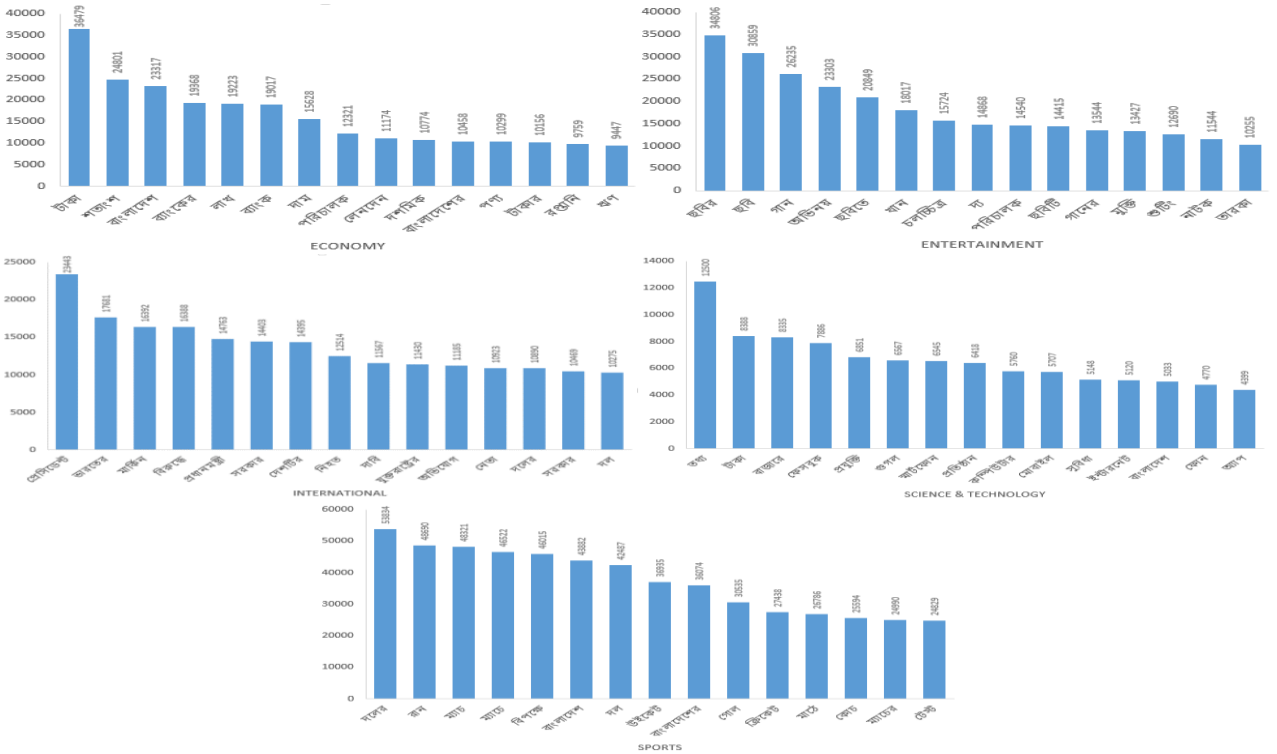


Fig. 4. Fourteen most Frequent Words of each Category after Data Cleaning of Dataset II.

TABLE VI. TF-IDF MATRIX OF FIRST TEN DOCS AND FIRST SIX WORDS OF DATASET I

	WORD 1	WORD 2	WORD 3	WORD 4	WORD 5	WORD 6
DOC 1	0	0	0	0	0	0
DOC 2	0	0	0	0	0	0
DOC 3	0	0	0.038	0	0	0
DOC 4	0	0	0	0	0	0
DOC 5	0	0.038	0	0	0	0
DOC 6	0	0	0	0	0	0
DOC 7	0.032	0.033	0	0	0	0
DOC 8	0	0	0	0	0	0
DOC 9	0	0	0	0	0	0
DOC10	0	0	0	0	0	0

E. Splitting Dataset into Training and Testing Set

After Successful feature extraction, Datasets are split into train and test datasets. Both datasets are divided into 4:1. Four portions of dataset are used for the training set, and the rest portion is for testing. That means 80% of data from the datasets are used for training, and the rest 20% is considered as the testing. This step is done using the sklearn library, which is very simple.

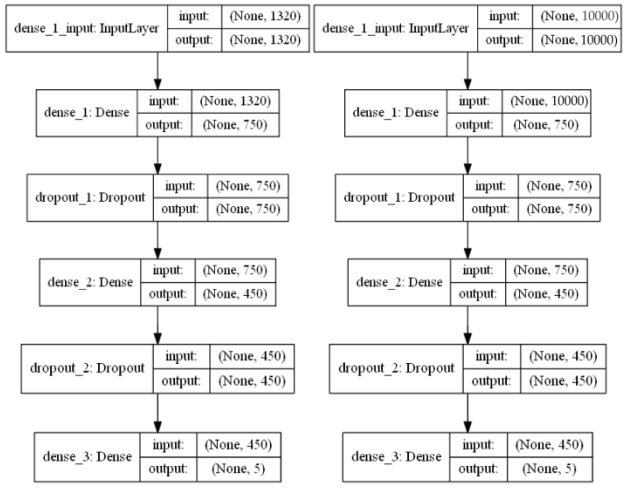
F. Building and Fitting Models

In this stage, the datasets are fitted into different machine learning classifier algorithms and neural network.

1) Using machine learning classifier algorithm: Here several machine learning classifiers used such as Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, Decision Tree for the classification of Bangla news & import sklearn built-in classifier for this.

2) Using multi-layer dense neural network: Here the data preprocessing technique and feature extraction technique is the same as machine learning algorithms. However, one hot encoder is used for encoding the encoding category. For categorizing task, feed forward neural network is used as classification algorithm. It is organized in the form of multiple layers. In the proposed model, dense layer has been used. Feed forward neural network consists of the input layer, the hidden layers and the output layer. Dataset generated Input patterns are transmitted from input layer to next layer which is also called by first hidden layer. Later output from the first hidden layer is being used as the input of the second hidden layer. The same process continueing untill reach the last hidden layer. Finally, the output of the last hidden layer is being used as the input of output layer or last layer. For building such model, Sequential model has been used. This model uses a linear stack of layers. The most common layer is a dense layer which is a regular densely connected neural network layer with all the weights and biases. In the first layer input shape is determined since the following layers can make automatic shape inference. To build the Sequential model, layers one by one are added in order. Total three dense layers used. After each dense layer, one dropout layer added with a 20% dropout rate.

For the layers between the input and output layer, relu activation function is used, and in the output layer, the softmax function is used as an activation function. Before training the model, the learning process is configured using optimizer and loss function. Here adam optimizer and categorical_crossentropy is used as optimizer and loss function respectively to train the model. Different numbers of dense layers, different numbers of nodes between layers, and different epochs are used for checking the accuracy of the model. Finally, the model is built by adding three dense layers. On the first layer, 750 nodes are added, on the second layer 450 nodes, and on the third layer, 5 nodes are added as the dataset is of five categories. The multi-layer dense neural network model for the two datasets is depicted in Fig. 5. The model is trained in 200 epochs. By doing so, the highest accuracy is got for both Datasets. Step by step procedure that is done for building a multilayer dense neural network model for getting the highest accuracy Bangla news classification is represented in Fig. 6. How the training loss and accuracy changes in both Dataset is represented in Fig. 7 and Fig. 8, respectively.



(a) For dataset I (b) For dataset II

Fig. 5. Multi-layer Dense Neural Network Model.

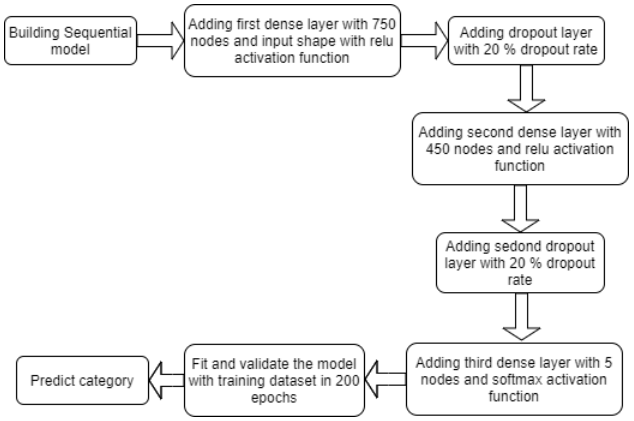


Fig. 6. Step by Step Procedure of Feed-forward Multi-layer Dense Neural Network for Bangla News Classification.

Fig. 7. Training and Validation Accuracy and Loss for Dataset I.

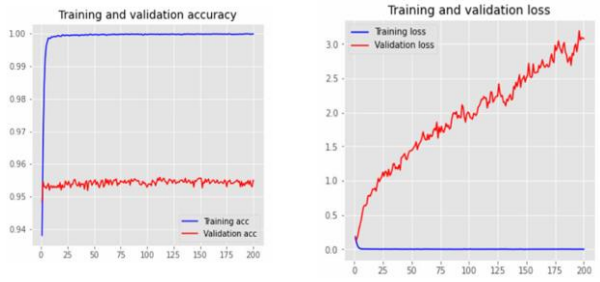


Fig. 8. Training and Validation Accuracy and Loss for Dataset II.

G. Category Prediction

After fitting Dataset to classifier, the main job is to predict category. In this stage, the model is trained to predict the test data that are unseen to the machine. If any vectorized sample of Bangla news is given to the model, it can predict the category of the sample data. Here for training, as different classifiers are used such as Naïve Byes, Decision Tree, K-Nearest Neighbors, Support Vector Machine, Random Forest, and also the model which is made by neural network can predict the category of vectorized sample data.

The model and vectorizer are saved as a pickle file and then it is used to classify. A random sample is given to the model and it successfully classifies the random sample. A screenshot of the random testing is given in Fig. 9.

```
sample
```

```
['১৯ বছর ধরে খেলছেন বাংলাদেশ দলে। জুলাইতে ২৭০ ওয়ানডে উইকেট। এক সময়ের অভিজ্ঞতা। কেবল পারফরম্যান্সের বিচারেই মাশরাফি বিন মুর্তজার থাকেনি কোনো বাংলাদেশি পেসার। কিন্তু মাশরাফিকে পারফরম্যান্স দিয়ে বিচার করা যাবে কি। নেতা মাশরাফিও তো নিজেকে অন্য উচ্চতায় নিয়ে যাওয়া কেউ। দুটি বিশ্বকাপে দেশকে নেতৃত্ব দিয়েছেন। অসাধারণ প্রেরণাদায়ী ব্যক্তিত্ব। দলে তাঁর উপস্থিতিই বিরাট ব্যাপার। বাংলাদেশ দলে তো এমনটা ছিলেনই। ঘরোয়া ক্রিকেট বা যেকোনো ফ্র্যাঞ্চাইজি লিগেও যে দলকে নেতৃত্ব দিয়েছেন, যে দলে খেলেছেন, সেটিকেই নিজের প্রেরণার জাদুতে বদলে দিয়েছেন। তাঁর অনুপস্থিতি অনুভবটা দীর্ঘই হবে বাংলাদেশের ক্রিকেটে। \n\n ‘অনুপস্থিতি’ শব্দটা লেখা হচ্ছে বটে, কিন্তু মাশরাফি তো এখনো খেলছেন। এ শব্দটি লিখতে হচ্ছে কারণ তাঁকে সদ্যই জাতীয় দল থেকে বাদ দেওয়া হয়েছে। আসছে ওয়েস্ট ইন্ডিজ সিরিজের জন্য যে ২৪ সদস্যের প্রাথমিক দল ঘোষণা করা হয়েছে, তাতে নেই জাতীয় দলের সাবেক এ অধিনায়ক। নির্বাচকেরা বলছেন, কঠিন একটা সিদ্ধান্তই তাঁদের জন্য এই মাশরাফিকে বাদ দেওয়া। কিন্তু সেটি করতে হয়েছে বাংলাদেশ ক্রিকেটের ভবিষ্যতের কথা চিন্তা করে। বয়স ৩৭ হয়ে গেছে। ক্যারিয়ারের সায়াহ্নে মাশরাফি পৌঁছে গেছেন এমনিতেই। ২০২৩ বিশ্বকাপের আগে দলের নতুন রক্ত ঢোকাতে হবে–মাশরাফিকে তাঁরা বাদ দিয়েছেন সেসব কথা চিন্তা করেই। ']
```

```
sample=tfidf.transform(sample).toarray()
classs=clf.predict(sample)
if  classs==0:
    print(' It is echonomical news')
elif classs==1:
    print(' It is entertainment news')
elif classs==2:
    print(' It is international news')
elif classs==3:
    print(' It is scn_and_tech news')
else:
    print(' It is sports news')
print(clf.predict(sample))
```

```
 It is sports news
[4]
```

Fig. 9. Screenshot of Output of Successfully Classifying given Sample.

IV. Result and Discussion

In this section, the performances of the model are analyzed on different machine learning algorithms and neural network for both Datasets.

A. Accuracy of Model

Onfusion matrix is a presentation for summarizing the performance of a classification algorithm. The confusion matrix for all classifier algorithms is given in Fig. 10 and Fig. 11 for Dataset I and Dataset II, respectively. By judging the confusion matrix the best model can be decided. From this matrix, the accuracy, precision, recall, and f1-score of the built model can be calculated. For different classifiers, different confusion matrices are built and from those confusion matrices, the accuracy, precision, recall, and f1-score of different classifiers is calculated. The performances of the different models are represented in Tables VII and VIII for dataset I and dataset II respectively and the highest precision, recall and accuracy category wise for all the classifiers are shown as bold font. Overall performance of different classifier model is shown in Tables IX and X for both dataset respectively. Here, highest accuracy, precision and recall, and f1-score of algorithms are shown also as bold font. Fig. 12 is a plot of the accuracy of different classifiers for both datasets. F1-score of classifiers according to news type is shown on Fig. 13 and Fig. 14 for Dataset I and Dataset II respectively.

1) Naïve bayes: Both dataset have five categories. From Tables VII and VIII it is clear that there are variations in different performance rate for different types of news. For dataset I Entertainment has the lowest f1-score and for dataset II sports has the lowest f1-score when Naïve Bayes classifier is used. If all the categories are combined the accuracy for Naïve Bayes model is 91.23% and 92.76% for dataset I and dataset II, respectively.

2) K-Nearest neighbor: The accuracy of this model is 84.81 % for dataset I and 70.4% for dataset II. The lowest performance rate is found for entertainment category in dataset I and the same is international category in dataset II.

3) Support vector machine: Support Vector Machine can be defined over a vector space where the problem is for finding a decision surface that “best” separates the data points in two classes [9]. Support Vector Machine has different types of kernels. The linear kernel is used for the purpose. Overall accuracy of this model is 89.12% for dataset I and 94.99% for dataset II. Here science and technology category shows the lowest performance rate for dataset I and for dataset II sports category shows the lowest performance.

4) Random forest: For building Random Forest classifier model entropy criterion and 50 n_estimators are used here and got accuracy 87.01% for dataset I and 91.4% for dataset II. In the prediction model the lowest rate of performance is found in science and technology category for dataset I and sports for dataset II.

5) Decision tree: The overall accuracy of this model is 62.45% for dataset I and for dataset II accuracy is 79.87%. Again for dataset I science and technology has the lowest rate of performance and for dataset II sports has the lowest rate of

performance in this prediction model. On the other hand the highest rate of performance is found in sports category for dataset I and science and technology is for dataset II.

6) Logistic regression: Again, the lowest accuracy is found for science and technology category in dataset I and sports category in dataset II in this model.. The highest rate of performance in the prediction model is sports for dataset I and science and technology for dataset II. The overall accuracy of this model is 90.52 % for dataset I and 94.6% for dataset II.

7) SGD classifier: If all the categories are combined to get the accuracy of the SGD classifier, the accuracy is 88.77% and 93.78% for dataset I and dataset II, respectively. The SGD confusion matrix is shown in Fig. 10(g) and 11(g) for both dataset which shows which category has high performance. For dataset I sports has the highest f1-score and science and technology category has the lowest f1-score and for dataset II science and technology has the highest f1-score and sports has the lowest f1-score.

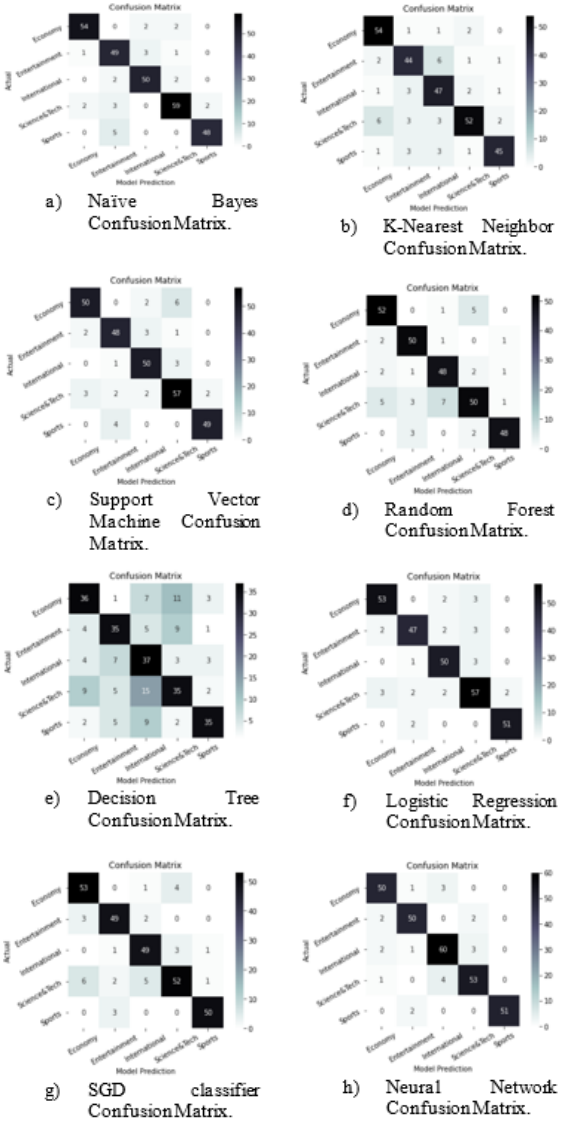


Fig. 10. Confusion Matrix of different Classifiers for Dataset I.

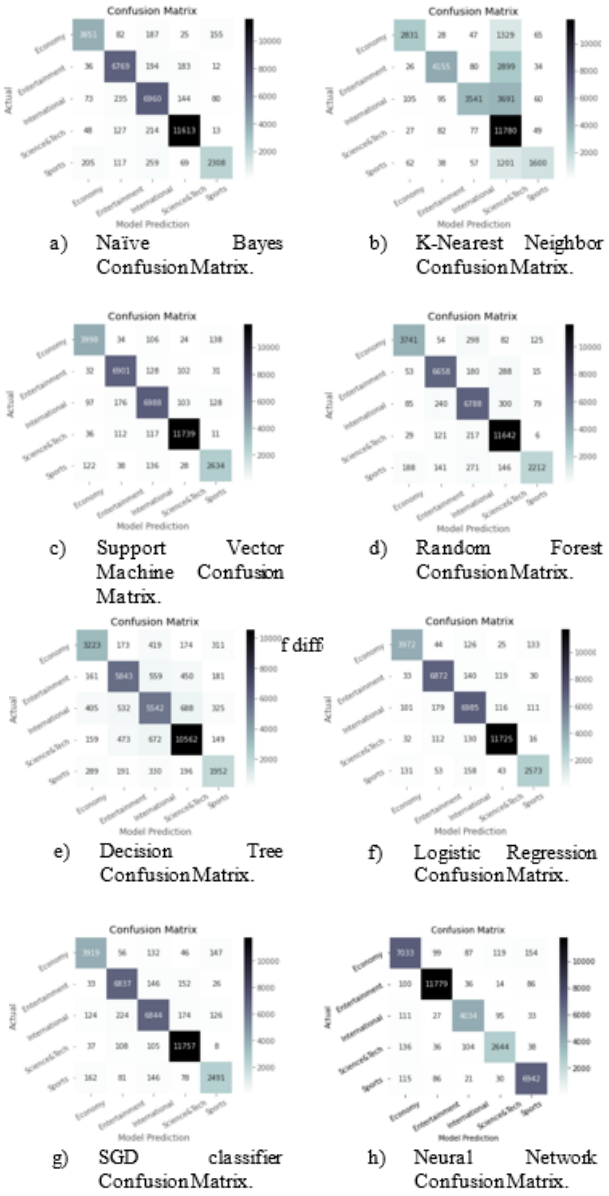


Fig. 11. Confusion Matrix of different Classifiers for Dataset II.

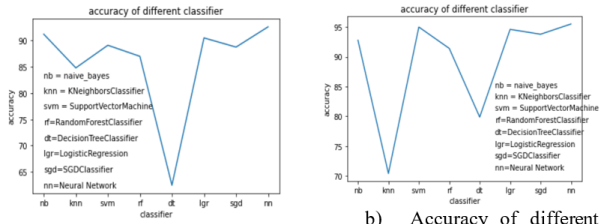
TABLE VII. RESULT COMPARISON BETWEEN MACHINE ALGORITHMS OF DATASET I

Classifiers	Category	Precision	Recall	F1-score
Naïve Bayes	Economics	0.94	**0.93**	**0.93**
	Entertainment	0.83	0.90	0.86
	International	0.91	0.92	0.91
	Science&Tech	0.92	0.89	0.90
	Sports	**0.96**	0.90	**0.93**
K-Nearest Neighbor	Economics	0.84	**0.93**	**0.88**
	Entertainment	0.81	0.81	0.81
	International	0.78	0.87	0.82
	Science&Tech	0.89	0.78	0.83
	Sports	**0.91**	0.85	0.87

Support Vector Machine	Economics	0.90	0.86	0.88
	Entertainment	0.87	0.88	0.87
	International	0.87	**0.92**	0.89
	Science&Tech	0.85	0.86	0.85
	Sports	**0.96**	**0.92**	**0.94**
Random Forest	Economics	0.85	0.89	0.87
	Entertainment	0.87	**0.92**	0.89
	International	0.84	0.89	0.86
	Science&Tech	0.84	0.75	0.79
	Sports	**0.94**	0.90	**0.92**
Decision Tree	Economics	0.65	0.62	0.63
	Entertainment	0.66	0.65	0.65
	International	0.51	**0.68**	0.58
	Science&Tech	0.58	0.53	0.55
	Sports	**0.79**	0.66	**0.72**
Logistic Regression	Economics	0.91	0.91	0.91
	Entertainment	0.90	0.87	0.88
	International	0.89	0.92	0.90
	Science&Tech	0.86	0.86	0.86
	Sports	**0.96**	**0.96**	**0.96**
SGD Classifier	Economics	0.85	0.91	0.88
	Entertainment	0.89	0.90	0.89
	International	0.85	0.90	0.87
	Science&Tech	0.88	0.78	0.82
	Sports	**0.96**	**0.94**	**0.95**
Multi-layer Dense Neural Network	Economics	0.91	0.93	0.92
	Entertainment	0.93	0.93	0.93
	International	0.90	0.91	0.90
	Science&Tech	0.91	0.91	0.91
	Sports	**1.00**	**0.96**	**0.98**

TABLE VIII. RESULT COMPARISON BETWEEN MACHINE LEARNING ALGORITHMS OF DATASET II

Classifiers	**Category**	**Precision**	**Recall**	**F1-score**
Naïve Bayes	Economics	0.91	0.89	0.90
	Entertainment	0.92	0.94	0.93
	International	0.89	0.93	0.91
	Science&Tech	**0.96**	**0.96**	**0.96**
	Sports	0.90	0.78	0.83
K-Nearest Neighbor	Economics	0.92	0.66	**0.77**
	Entertainment	**0.94**	0.57	0.71
	International	0.93	0.47	0.62
	Science&Tech	0.56	**0.98**	0.71
	Sports	0.88	0.54	0.67
Support Vector Machine	Economics	0.93	0.93	0.93
	Entertainment	0.95	0.96	0.95
	International	0.93	0.93	0.93
	Science&Tech	**0.97**	**0.97**	**0.97**
	Sports	0.89	0.89	0.89
Random Forest	Economics	0.91	0.87	0.89
	Entertainment	0.92	0.92	0.92
	International	0.87	0.90	0.88
	Science&Tech	**0.93**	**0.97**	**0.95**
	Sports	0.90	0.74	0.81
Decision Tree	Economics	0.76	0.75	0.75
	Entertainment	0.81	0.81	0.81
	International	0.73	0.74	0.73
	Science&Tech	**0.87**	**0.88**	**0.87**
	Sports	0.67	0.66	0.66
Logistic Regression	Economics	0.93	0.92	0.92
	Entertainment	0.94	0.95	0.94
	International	0.92	0.93	0.92
	Science&Tech	**0.97**	**0.97**	**0.97**
	Sports	0.90	0.87	0.88
SGD Classifier	Economics	0.91	0.91	0.91
	Entertainment	0.93	0.95	0.94
	International	0.93	0.91	0.92
	Science&Tech	**0.96**	**0.98**	**0.97**
	Sports	0.89	0.84	0.86
Multi-layer Dense Neural Network	Economics	0.94	0.94	0.94
	Entertainment	**0.98**	**0.98**	**0.98**
	International	0.94	0.94	0.94
	Science&Tech	0.91	0.89	0.90
	Sports	0.96	0.96	0.96



a) Accuracy of different classifier for Dataset I

b) Accuracy of different classifier for Dataset II

Fig. 12. Accuracy of different Classifier for both Dataset.

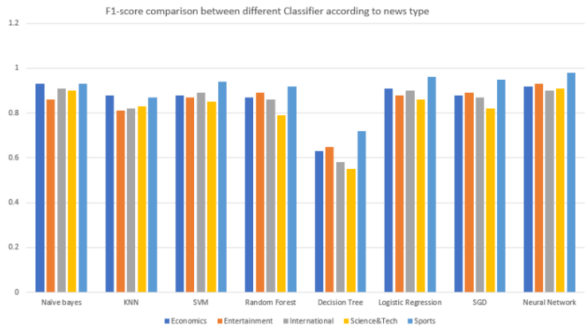


Fig. 13. Comparison of f1-score of Dataset I of different Classifier according to the News Types.

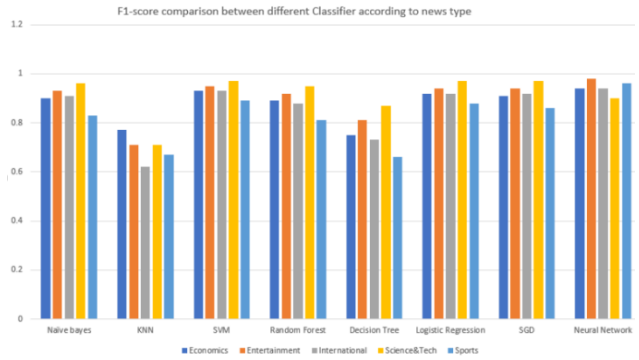


Fig. 14. Comparison of f1-score of Dataset II of different Classifier according to the News Types.

8) Multi-layer dense neural network: In this case, international news for dataset I and science and technology news for dataset II have the lowest rate of performance. In Machine learning model it can be seen that, In dataset I, all classifiers returned lowest performance in science & Technology category, except Naive Bayes and K-Nearest Neighbor. Where in dataset II, all classifiers gives lowest performance in sports category where only K-nearest neighbour gives lowest performance on other category. But ,it doesn't happen in multi-layer dense neural network model. The accuracy of this model is also quite impressive. For dataset I sports and for dataset II, entertainment has the highest rate of performance. This model has the highest accuracy comparing the other traditional machine learning models. The overall accuracy of this model is 92.63% for dataset I and 95.50% for dataset II.

B. Comparison of Algorithms

Table IX shows that in the traditional machine learning algorithms for dataset I the highest result comes from the Naïve Bayes classifier model and Table X shows for dataset II Support Vector Machine has the highest result. But for both dataset the highest accuracy comes from multi-layer dense neural network. That means multi-layer dense neural network gives the best performance for both dataset. In Table IX it is also shown that decision tree classifier gives the worst result for dataset I. On the other hand from Table X, it is shown that k-nearest neighbor classifier gives the worst result. If the confusion matrix of Fig. 11(b) is observed it is seen that the highest false classification is found for k-nearest neighbor. From Tables IX and X of overall performance, it is shown that most of the classifiers have low variance and low bias which indicates the proposed model doesn't have underfitting and overfitting.

TABLE IX. OVERALL PERFORMANCE OF DIFFERENT CLASSIFIER MODEL ON DATASET I

Classifiers		**Accuracy**	**Precision**	**Recall**	**F1-score**
Type	**Name**				
Machine Learning Algorithm	Naïve Bayes	**91.23%**	**91.37%**	**91.28%**	**91.32%**
	K-Nearest Neighbor	84.81%	85.13%	85.06%	85.09%
	Support Vector Machine	89.12%	89.41%	89.30%	89.35%
	Random Forest	87.01%	87.20%	87.49%	87.34%
	Decision Tree	62.45%	64.01%	62.29%	63.44%
	Logistic Regression	90.52%	90.72%	90.71%	90.72%
	SGD Classifier	88.77%	88.96%	89.19%	89.08%
Neural Network	Multi-layer Dense Neural Network	**92.63%**	**93%**	**92.8%**	**92.8%**

TABLE X. OVERALL PERFORMANCE OF DIFFERENT CLASSIFIER MODEL ON DATASET II

Classifiers		**Accuracy**	**Precision**	**Recall**	**F1-score**
Type	**Name**				
Machine Learning Algorithm	Naïve Bayes	92.76%	91.84%	90.24%	91.03%
	K-Nearest Neighbor	70.4%	85.05%	64.6%	73.42%
	Support Vector Machine	**94.99%**	**93.84%**	**93.78%**	**93.81%**
	Random Forest	91.4%	91.07%	88.36%	89.7%
	Decision Tree	79.87%	77.03%	76.81%	76.92%
	Logistic Regression	94.6%	93.54%	93.14%	93.34%
	SGD Classifier	93.78%	92.68%	91.92%	92.30%
Neural Network	Multi-layer Dense Neural Network	**95.50%**	**94.6%**	**94.2%**	**94.4%**

V. Conclusion and Future Work

The main focus of this research is to build an automatic classification system for Bangla News documents. This system provides users an efficient and reliable access to classified news from different sources. Different as well as most widely used machine learning classifiers and multi-layer dense neural network are used for categorization and a comparison has been conducted between them. Among the classifier algorithms, Support Vector machine Classifier provides the best result. In the model, TF-IDF technique is used for vectorization to fit data to the classifier.

In future, word2vec model will be used for better result and for preventing the limitation of TF-IDF model. In TF-IDF model, more importance is put on the uncommon words. But, semantic information of the words is not stored in TF-IDF model.

In this research, multi-layer dense neural network and some built in classifier like Naïve Bayes classifier, k-nearest neighbor classifier, random forest classifier, support vector machine classifier and decision tree classifier were used. In future, CNN, RNN and other neural network model will be examined to build the model for better performance.

References

[1] Tenenboim, L., Shapira, B. and Shoval, P. 2008. “Ontology-based classification of news in an electronic newspaper”.

[2] Pendharkar, B., Ambekar, P., Godbole, P., Joshi, S. and Abhyankar, S. 2007. “Topic categorization of rss news feeds, Group”.

[3] Carreira, R., Crato, J. M., Gonçalves, D. and Jorge, J. A. 2004. “Evaluating adaptive user profiles for news classification”. 9th international conference on Intelligent user interfaces, pp. 206–212.

[4] Fauzi, M. A., Arifin, A. Z., Gosaria, S. C., & Prabowo, I. S. (2016). “Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model”. Indonesian Journal of Electrical Engineering and Computer Science, 2(3), 401-408.

[5] Dutta, K., Kaushik, S., & Prakash, N. (2011). “Machine learning approach for the classification of demonstrative pronouns for Indirect Anaphora in Hindi News Items”. The Prague Bulletin of Mathematical Linguistics, 95(1), 33-50.

[6] El-Barbary, O. G. (2016). “Arabic news classification using field association words”. Advances in research, 1-9.

[7] Beresi, U. C., Adeva, J. G., Calvo, R. A., & Ceccatto, A. H. (2004, August). “Automatic classification of news articles in Spanish”. In Actas del Congreso Argentino de Ciencias de Computación (CACIC) (pp. 1588-1600).

[8] Kabir, F., Siddique, S., Kotwal, M. R. A., & Huda, M. N. (2015, March). “Bangla text document categorization using stochastic gradient descent (sgd) classifier”. In 2015 International Conference on Cognitive Computing and Information Processing (CCIP) (pp. 14). IEEE.

[9] Y. Yang and X. Liu, “A re-examination of text categorization methods,” in Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ser. SIGIR ’99. New York, NY, USA: ACM, 1999, pp. 42–49. [Online]. Available: http://doi.acm.org/10.1145/312624.312647.

[10] Chy, Abu Nowshed & Seddiqui, Hanif & Das, Sowmitra. (2014). “Bangla news classification using naive Bayes classifier”. 16th Int'l Conf. Computer and Information Technology, ICCIT 2013. 10.1109/ICCITechn.2014.6997369.

[11] FouziHarrag,Farhat ABBAS University,Eyas EI Qawasmah,JUST University,“Neural Network for Arabic Text Classification”, in 2009 Second International Conference on the Applications of Digital Information and Web Technologies, doi:10.1109/ICADIWT.2009.5273841.

[12] Selamat, A., & Omatu, S. (2003, July). Neural networks for web page classification based on augmented PCA. In Proceedings of the International Joint Conference on Neural Networks, 2003. (Vol. 3, pp. 1792-1797). IEEE.

[13] Hossain, M. R., Sarkar, S., & Rahman, M. “Different Machine Learning based Approaches of Baseline and Deep Learning Models for Bengali News Categorization”. International Journal of Computer Applications, 975, 8887.

[14] Manisha Chakraborty, and Mohammad Nurul Huda, “Bangla Document Categorization using Multilayer Dense Neural Network with TF-IDF”, International Conference on Advances in Science, Engineering Robotics Technology (ICASERT 2019), May 3-5, 2019, Dhaka, Bangladesh, pp. 1-4.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[15] Mandal, A. K., & Sen, R. (2014). “Supervised learning methods for bangla web document categorization”. arXiv preprint arXiv:1410.2045.

[16] M. M. H. Shahin, T. Ahmmed, S. H. Piyal and M. Shopon, "Classification of Bangla News Articles Using Bidirectional Long Short Term Memory," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 1547-1551, doi: 10.1109/TENSYMP50017.2020.9230737.

[17] twintyOne, “Prothom alo [2013 - 2019]”, Aug 2019. [Online]. Available: https://www.kaggle.com/twintyone/prothomal .

[18] stopwords-iso stopwords-bn. [Online]. Available: https://github.com/stopwords-iso/stopwords-bn.

[19] Bengali Stopwords, [Online]. Available: https://www.ranks.nl/stopwords/bengali?fbclid=IwAR3gFJxN8Yo3BT6S9bMGY87NQzudqMf7z9kxX4veH0aYMLZBBrDaCrZH1jo.

[20] sharminyeasmin198, “Bengali News Dataset (Prothom alo)”, March 2021. [Online]. Available: https://www.kaggle.com/sharminyeasmin198/bengali-news-dataset-prothom-alo.

[21] sharminyeasmin198, “Bengali Stopwords 2021”, March 2021. [Online]. Available: https://www.kaggle.com/sharminyeasmin198/bengali-stopwords-2021.