A comparative study on Bangla news classification using machine learning approach

***Abstract*—News classification has been increasing popularity in recent years. With the increasing use of mobile applications, users can now access multimedia content, newspapers, magazines and other online resources. Scholars logged a massive quantity of unstructured information and explored the literature for different techniques of translating this scattered material into a single, organized volume due to the ubiquitous accessibility of text in diverse versions. In the long text, there is a greater emphasis on full classification (full information, massive data, extended message, etc) than in the short text. Newspaper stories cover a wide range of topics, including economics, sports, politics, businesses, entertainments, and more. The number of Bangladeshi online news sites on the internet has also increased. Online readers will benefit from being steered to their selected news category, which will help them discover desired material. It takes a long time and a lot of effort to identify news pieces manually. As a result, in today, where enormous amounts of unsorted text are a concern, text classification is critical. There has already been a lot of research on the classification of Bengali news. Most of the researchers applied SVM, Naive Bayes, Random Forest, KNN, RNN, BiGRU, CNN, SVM, or LSTM in their research work. The main purpose of this study is to evaluate and identify the best machine learning classifiers for Bangla news classification. By analyzing and comparing some recent research works and the proposed methods, we have found a hybrid model consisting of CNN, LSTM and glove vectorization that gives better accuracy.**

***Index Terms*—News Classification, NLP, Comparative, CNN, LSTM, GLoVe vectorization**

I. Introduction

News classification is one kind of text classification. Information retrieval and tracking, speech annotation, text classification, and other uses of Natural Language Processing (NLP) have gained in popularity in recent years. Text classification is particularly useful for a variety of purposes, including controlling online content, search engines, and web filtering. As technology advances, people are becoming more interested in text classification issues. There are a plethora of Bengali language publications available online that are both useful and difficult to categorize into their sub-sets. In addition, customers want to get a personalized version of the newspaper with relevant stories featured on front page. This type of job is done on several international news and blog sites. As a result, text classification is a commercial as well as a job with time-saving results.

Text classification is a strategy for categorizing documents using NLP. Due to the large size of feature vectors, which consist of irrelevant and unrelated data, text classification is a difficult process. Researchers have used this Text Classification to conduct a lot of work during the last few decades. The most typical tools for extracting important characteristics for classification are Word2Vec [1], TF-IDF [2], and Doc2Vec [3]. For text classification, it is also a good idea to utilize a variety of Supervised approaches, such as Naive Bayes [6], KNN [5], or SVM [4]. Researchers have also used unsupervised approaches for text classification tasks. I have found several BNLP and Bengali speech processing materials and methods to provide in this study. The elements of natural language processing are shown in Figure 1.

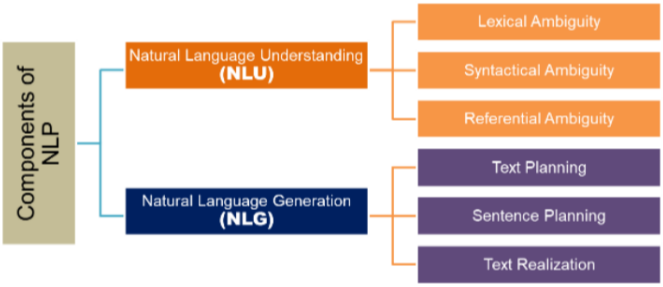

Fig. 1. Natural Language Processing Components [17]

In recent years, scholars have been more interested in machine learning and deep learning models like SVM, Naive Bayes, Random Forest, KNN, RNN, BiGRU, CNN, SVM, and LSTM. According to a paper on Bengali Text classification [5]. 9127 news pieces were divided into nine categories and the highest performing classifier is SGD. Despite Bengali's long history and status as among the world's most widely used languages, native Bengali speakers make only for about 8% of the global population [3]. organizing and categorizing Bengali content is important so that users may find relevant information.

II. Related Work

Text classification in Bengali has been available since the early 2000's. At first there were limited datasets to work with, but things have changed lately, and a significant number of datasets can be found in most areas where the NLP algorithm can be deployed.

According to Alam et al. [9], they have used a public dataset of Bengali newspapers in their research. The article is divided into ten sections and contains 84,906 Bengali text information. To classify Bengali text data, they have employed machine learning algorithms such as SVM, SGD, MNB, and Logistic Regression. Firstly, they divided the data-set into 10027, 42370, 60000, and 84906 sections and then classify them.

They discovered that accuracy decreased as the amount of training data increased. SVM for 84906 data sample, received a maximum of 87.5% accuracy for sample data.

N. Tabassum et al. [6] created a dataset that had 1,050 Bengali Twitter and Facebook comments by improving a sensitive experimental framework. The solutions include a unigram, POS tags, denial handling, and random forest classification in order to provide a more accurate outcome of roughly 85 percent.

Another researcher, Tej Bahadur Shahi, provided a prediction for the self-acting Nepali news multi-category [6]. She is also finished her study on neural networks and machine learning classifiers. Multi-layer connectivity is utilized with deep learning classifiers like Naive Bayes and Support Vector Machine. However, there are some minor details. The neural network is in an awkward predicament. During the procedure, Nepali newspaper text categorization was effective at 74.65% using SVM with RBF. However, with 73 percent accuracy, Neural Network comes in second. Data Volume of Nepali News Text Classification There were 4,964 in all, with 20 distinct sorts of news. In deep learning models in general Large amounts of data are required by such neural networks. M Rahman et al. [7] Bengali papers were categorized using BERT and ELECTRA transformers on three separate datasets, with satisfactory results. According to their findings, BERT has fallen behind Electra in terms of overall dataset correctness. Furthermore, ELECTRA outscored BERT in most circumstances, with top scores of 94.18 percent and 96.39 percent in most cases.

Another study [2] conducted by H. Berger et al. examined text categorization algorithms employing N-gram frequency statistics. In addition, three classification methods were used to N-gram character statistical features and word density document representation (PART, SMO, and NBM). They used a multilingual electronic communications dataset consisting of 1,811 emails that had been manually divided into layers. In addition, the impact of information on classifier performance was investigated. In SMO-related emails, their research proved the efficacy of multi-class email categorization.

A. Sharfuddin et al. [4] used a DRNN with BiLSTM implementation to present a solution for sentiment analysis. Their dataset consisted of 10,000 Facebook comments, 5,000 of which were positive and the rest were unfavorable. The dataset was trained using their recommended approach, BiLSTM, which achieved an accuracy of 85.67%. Meanwhile, they used SVM, Decision Tree Classifier, and Logistic Linear Regression to train the dataset, attaining accuracy of 68.77 percent, 67.50 percent, and 60.94 percent, respectively.

For Bengali Document Categorization, a companion article [5] compares Naive Bayes, SGD, and SVM. Many feature selection approaches, such as normalized TFIDF and also Chi-square distribution using a word analysis system, were employed to evaluate the above-mentioned classifier's usefulness in predicting a document category. SVM classifier got the highest F1-score of 92.56 percent after employing numerous techniques, while normalized TFIDF was utilized to select features and CHI-square was coupled with NB to achieve the lowest F1-score of 83.36 percent.

III. News Classification

News classification is a type of text categorization. Text categorization are used to sort natural texts into classifications. Consider categorizing news by subject or categorizing book reviews by negatively or positively feedback. Text categorization is also useful for detecting languages, managing client feedback, and fraud detection. NLP based model can simplify this procedure, which is time-consuming when performed manually.

For news, categorization is a multi-label text categorization issue. The purpose is to allocate a news item with one maybe more category. A series of classification algorithm is a common strategy for multi-label text categorization. Before releasing a news, each online news site categories it so that viewers may quickly find the sort of coverage that fascinates them the next time they appear. As example, I enjoy reading about the newest technological developments, therefore I always go to the technology section of a news site. However, people may or not be willing to read about technologies; instead, users could be curious, commerce, culture, or sport activities. Usually, news items are categorised by manually by web page administrators. They can save time by implementing any natural language processing model on their sites that reads the title or text and organizes the news by classification.

IV. Methodology

The research is focused on Bangali news classification using machine learning algorithms. Most of the researchers used CNN, Naive Bayes, LSTM, BiGRU, SVM, Logistic Regression, and KNN algorithms in their proposed method.

To fix the Bengla news sentiment categorization problem, Islam et al. [7] used a variety of Machine Learning and Natural Language Processing algorithms. Naive Bayes and SVM provide better results than other algorithms in their research. Mohiuddin, & Matin [8] build their model based on a Recurrent Neural network with BiGRU.

Using CNN and Bi-LSTM, Alam et al. [9] proposed their model.

Chowdhury et al. [10] applied CNN and LSTM with GloVe in their proposed model. They claim that their suggested model is more efficient since it performs with ten categories, preprocessed data utilizing 'GloVe' and filter-based analogies employing sophisticated LSTM features.

LibSVM were recommended used by Hasan et al. [11] in the model. They train their system in 2 libraries. The first one is with liSVM and the second one was with Scikit learn.

Shuvo et al. [12] applied CNN in their model. A Neural Network based model were suggested by Khushbu et al. [13]. In their model, they utilized the GRU and LSTM algorithms and get better performance in GRU based model.

Dhar et al. [15] recommended creating a hybrid model using both TFIDF and counter vectorizer for this work. Ahmed et

al. [16] proposed their model based on Machine learning algorithms. They used both the SVM and RBF algorithms in their model. To run the data set in proposed model, Researchers need to prepossess the collected data-set. The processes that given below are followed by most of the researchers.

A. Data cleaning

Data cleaning is the first step of data prepossessing. In Bangla text there are a lot of noises. Because of this noise, interferes may occur in feature extraction. So researchers must remove it all from the data and keep as many of the key data as feasible. They have done some basic noise-reduction measures too.

1) Remove Stop-word: Stopwords are the most used words in a language. They are regularly removed from the text because they are irrelevant to the context and do not support feature extraction. The word stop can also be found in Bengali. There is no collection of benchmark stop words because it is a low-resource language. So, the researchers checked manually and remove them.

Some stop words example are shown in the fig 2

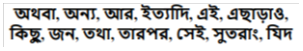

Fig. 2. Stop Words [9]

2) Remove Punctuation: For preparing data-set, all of the punctuation are needed to remove the selected data set because punctuation has no need in feature extraction. In Bangla, there are various type of punctuation. Also some of the punctuation has more than one uni-code. Regex are best and recommended by researchers for removing punctuation from the Bengali data-set.

3) Remove Unnecessary character: There exists many characters which are completely unnecessary and also they are not punctuation. So researchers also need to remove them from their dataset.

Some Unnecessary character's example are shown in the figure 3

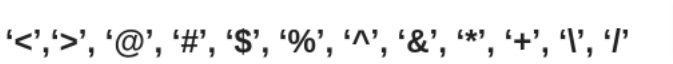

Fig. 3. Unnecessary character

4) Stemming: NLP relies heavily on data stemming. Stemming is the process of reducing a word to its base form. It decreases the dimension of the data. Some decent stemmers are available for Bengali because it is a low-resource language. Fig 4 is an example of the Bangla stemming procedure.

When data prepossessing is done, then the data-set are ready to run in model. Original and processed data are shown in Fig 5,

Analyzing researchers research methodology and their proposed model, a General Model for Bangla news classification using a machine learning approach is given below:

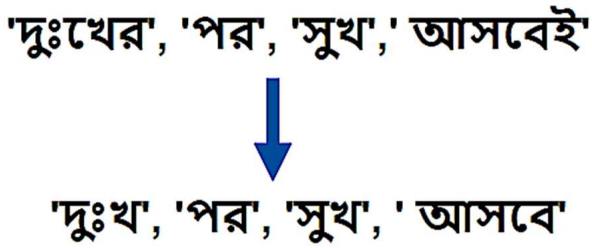

Fig. 4. Example of Bangla steaming processes

Original	Cleaned	Category
এবার হুমকি অধ্যাপককে। বন্ধু পুলিশকর্তাও নামলেন আক্রমণে	এবার হুমকি অধ্যাপককে বন্ধু পুলিশকর্তাও নামলেন আক্রমণে	state
কালো পতাকার বিক্ষোভের 'জবাব', প্রেসির জন্য ১১৮ কোটি টাকা বরাদ্দ করলেন মুখ্যমন্ত্রী	কালো পতাকার বিক্ষোভের জবাব প্রেসির জন্য ১১৮ কোটি টাকা বরাদ্দ করলেন মুখ্যমন্ত্রী	kolkata
সোশ্যাল মিডিয়ায় আক্রান্ত জাহ্নবী, কেন ট্রোলড হতে হল	সোশ্যাল মিডিয়ায় আক্রান্ত জাহ্নবী কেন ট্রোলড হতে হল	entertainment
বেজিং-নীতিকে আক্রমণ কংগ্রেসের	বেজিং নীতিকে আক্রমণ কংগ্রেসের	national

Fig. 5. Example of Original and processed data

B. Naive Bayes

Randomized machine learning based on the Bayes Theorem Naive Bayes may be applied to a variety of classification issues [7]. Positive and negative sentences are initiated and divided by the total amount of sentences in this method. Using Bayes Theorem [7], the statements are converted to words and categorized as positive or negative sentences.

C. SVM

The learned function of the SVM classifier differentiates positive and negative groups. Let us imagine the data points are X = x1, x2, x3..., and the weights are W= w1, w2, w3... The data items with weights are divided into two groups by

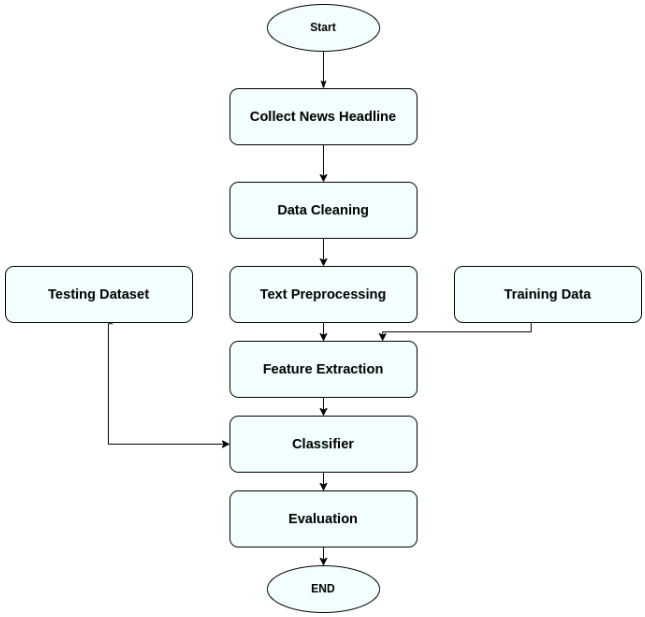

Fig. 6. Proposed Methodology

the Maximal-Margin hyperplane. The Lagrangian formulation may be used to define the Maximal-Margin hyperplane [7].

$$D(X^T) = \sum_{i=1}^{l} y_i \alpha_i X_i X^T + b$$

Here, X^T is a tuple, l denotes the support vector's number, and the class label X_i is designate by y_i. α_i and b are real number parameters that are calculated dynamically by the SVM algorithm.

D. GRU

A gated recurrent unit, shortly known as GRU, is a kind of RNN that utilizes connections between nodes to perform supervised learning like remembering and grouping in voice recognition. In order to prevent the vanishing gradient problem that plagues recurrent neural networks, GRU aid in the modification of neural network input weights [13].

E. BiGRU

A two-way GRU, often known as a BiGRU, is a sequential processing model that only has two GRUs. One takes the input towards the front, the other takes it backward. It does not need an explicit memory unit and instead uses reset and update gates. In comparison to LSTM, this requires less trainable parameters. The multilayer perceptron of this effectively predicts the pre trained Fast Text with such a 1D single hidden layer of 0.2 in Keras implementation.

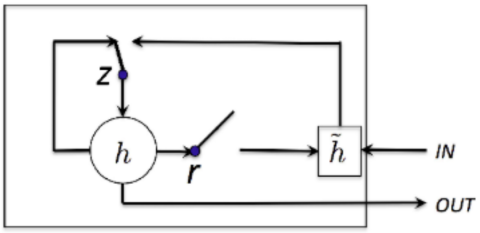

Fig. 7. Example of Original and processed data

F. CNN

CNN is a kind of neural network used for data analysis [9]. CNN is distinguished from other types of neural networks in that it analyses input as a 2D array rather than extracting attributes [9].

G. LSTM

Recurrent Neural Network topologies called LSTM networks are used to recall a short history of positive values [13]. It has three gates: an input gate that reads the input, an output gate that writes the output in the next stage, and the last one are forget gate that picks which input to remember and also which data to forget [13].

TABLE I
BANGLA NEWS CLASSIFICATION RESULT

Paper Author	Used Algorithm	Accuracy
Islam et al. [7]	Naive Bayes	73%
Islam et al. [7]	Support Vector Machine	75%
Mohiuddin et al. [8]	Bidirectional GRU	84.4%
Alam et al. [9]	CNN and Bi-LSTM	84.93%
Chowdhury et al. [10]	CNN-LSTM with GloVe vectorization	98.75%
Hasan et al. [11]	LibSVM	62.42%
Shuvo et al. [12]	CNN	93.65%
Hossain et al. [14]	SVM	43%
Khushbu et al. [13]	GRU	87.48%
Khushbu et al. [13]	LSTM	82.74%
Dhar et al. [15]	TFIDF with counter vectorizer	81%
Ahmed et al. [16]	SVM-RBF	63.71%

H. Bi-LSTM

Bi-LSTM, a type of RNN that enables any neural network to preserve sequential data in both backward and forward directions. In contrast to forward with and backward hidden sequences, a BiLSTM analyzes the input data from the reverse direction. The encoded vector is created by concatenating the final forward-backward output, which is the very first concealed layer's output sequences [9].

I. GloVe

Global vectors are employed in visual words and are referred to as GloVe [10]. It is just an unsupervised learning system created at Stanford for generating word sets from a corpus' global word-word co-occurrence matrix [10]. The result embedding in vector space reveal fascinating linear substructures of the word [10].

V. RESULT AND ANALYSIS

According to table 1, the authors applied the different data prepossessing and machine learning approaches and completed their work. Analyzing their proposed method and output, Chowdhury et al. [10] have the highest accuracy rate(98.75%).

Their proposed model was tested on a 14k data set that included Bangla news items from three different Indian newspapers in ten different sections. The current study in various fields is based on a statistical set consisting of sports news, Bangla comments, news of politics and also bangla twitter data. Bangla reviews, texts, and other data are unauthorized and require a lot of refinement [2]. Again, identifying political news will develop problems with limited terminology [5]. By comparison, their proposed model is more efficient because it handles up to 10 pre-processed data using 'GloVe', [10] a filter-based analogy that uses sophisticated LSTM features.

VI. Conclusion

We have compared different researchers' suggested classification on Bangla news using various machine learning algorithms. In one research, they used the CNN-LSTM with GloVe vectorization to train the classifier, and their accuracy was 98.75% [10]. Many researchers reported that CNN, an acronym for Convolutionary Neural Network, has better accuracy than other machine learning algorithms. In our opinion, CNN and LSTM, are the best approaches for Bangla news classification among the presented methods.

References

[1] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and word2vec for text classification with semantic features," in 2015 IEEE14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC), 2015, pp. 136–140.

[2] Z. Yun-tao, G. Ling, and W. Yong-cheng, "An improved tf-idf approach for text classification," Journal of Zhejiang University-Science A, vol. 6, no. 1, pp. 49–55, 2005.

[3] S. SrirangamSridharan, M. Srivatsa, R. Ganti, and C. Simpkin, "Doc2img: A new approach to vectorization of documents," in 2018 21st International Conference on Information Fusion (FUSION), 2018,

[4] A. Basu, C. Walters, and M. Shepherd, "Support vector machines for text categorization," in 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the, 2003, pp. 7 pp.

[5] Jingzhong Wang and Xia Li, "An improved knn algorithm for text classification," in 2010 International Conference on Information, Networking and Automation (ICINA), vol. 2, 2010, pp. V2–436–V2–439.

[6] Tabassum, N., Khan, M.I.: Design an empirical framework for sentiment analysis from bangla text using machine learning. In: 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). pp. 1–5. IEEE (2019)

[7] Islam, M. M., Masum, A. K. M., Rabbani, M. G., Zannat, R., & Rahman, M. (2019, November). Performance Measurement of Multiple Supervised Learning Algorithms for Bengali News Headline Sentiment Classification. In 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART) (pp. 235-239). IEEE.

[8] Mohiuddin, E., & Matin, A. (2021, July). Multilevel Categorization of Bengali News Headlines using Bidirectional Gated Recurrent Unit. In 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI) (pp. 1-6). IEEE.

[9] Alam, S., Haque, M. A. U., & Rahman, A. Bengali Text Categorization Based on Deep Hybrid CNN–LSTM Network with Word Embedding.

[10] Chowdhury, P., Eumi, E. M., Sarkar, O., & Ahamed, M. (2022). Bangla News Classification Using GloVe Vectorization, LSTM, and CNN. In Proceedings of the International Conference on Big Data, IoT, and Machine Learning (pp. 723-731). Springer, Singapore.

[11] Hasan, M. N., Bhowmik, S., & Rahaman, M. M. (2017, December). Multi-label sentence classification using Bengali word embedding model. In 2017 3rd International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-6). IEEE.

[12] Shuvo, M. I. R., Shahriyar, S. A., & Akhand, M. A. H. (2019, September). Bangla numeral recognition from speech signal using convolutional neural network. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-4). IEEE.

[13] Khushbu, S. A., Masum, A. K. M., Abujar, S., & Hossain, S. A. (2020, July). Neural network based bengali news headline multi classification system: Selection of features describes comparative performance. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[14] Hossain, E., Chaudhary, N., Rifad, Z. H., & Hossain, B. M. (2020). Bangla-news-headlines-categorization. GitHub.

[15] Dhar, P., & Abedin, M. (2021). Bengali News Headline Categorization Using Optimized Machine Learning Pipeline. International Journal of Information Engineering & Electronic Business, 13(1).

[16] Ahmed, T., Paul, R. R., Alam, A., Hasan, M., & Rab, R. (2022). Bangla News Popularity Prediction Using Machine Learning Techniques. Raqeebir, Bangla News Popularity Prediction Using Machine Learning Techniques (April 28, 2022).

[17] "Natural language processing (nlp) simplified : A step-by-step guide."Accessed: 2021-03-21.