International Journal of Research in Advanced Engineering and Technology
www.newengineeringjournal.in
ISSN: 2455-0876
Received: 08-02-2022, Accepted: 23-02-2022, Published: 10-03-2022
Volume 8, Issue 1, 2022, Page No. 10-16

Deep learning Bangla text classification using recurrent neural network

Ahsan Habib, Asma Akter
Department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Dhaka, Bangladesh

Abstract
Recent decades multi-level Bangla text classification is very crucial task for Bangla newspaper portal to improve their recommendation system as well as manual labour to categorize their different types of document. In the field of natural language processing and text mining a very few work has been done due its limited resource. The goal of this paper to provide a standard solution to overcome this constraints. In this paper a large number of dataset which consists almost 400k Bangla newspaper articles as JSON data format have utilized. We are using supervised machine learning model which consists of Long Short Term Memory (LSTM) for data extraction and data cleaning preprocessing method for Convolutional Neural Network (CNN). These newspaper articles have been categorized into nine categories named Bangladesh, Opinion, Internatioal, Education, Economy, Technology, Sports, Life-Style and Entertainment. Finally we understand the findings obtained from the model presented by various researcher's and prove that our model is more reliable than the previous framework.

Keywords: Bangla text classification, deep learning, LSTM, CNN, word-embedding

Introduction

Recently classification research has become the most common and crucial aspect in nature language processing. In machine learning approach various types of classification such as multi-level, multi-class and binary classification. Nowadays based on text classification various types of application are discovered such as sentiment analysis, emotion analysis, customer queries, categorizing articles etc. In this paper we categorize the articles of different Bangla newspaper. Newspaper carry the news of country's economic, trade, ecommerce, sports etc. In order to train our model we are using Kaggle website resource named Bangla newspaper dataset which is firmly preferable for our solution.

During the last decades there are many statistical and machine learning approach have been initialized to accurately classify the textual document. Deep learning is connected machine learning algorithm like support vector machine (SVM), KNN, Decision tree, Random forest and many more other classification algorithm. Deep learning technique like convolutional neural network and long short term memory are initialized to categorize the articles

Long Short Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems [1]. This network refers a complex area of deep learning. LSTM has the capability to handle gradient problem faced by Recurrent neural network (RNN). According to several online sources, LSTM has improved Google's speech recognition, greatly improved machine translations on Google Translate and the answer of Amazon's Alexa [2].

Convolutional Neural Network is class of deep neural networks. The objective of CNN is to extract high level features. CNN are special types of neural network that distinguished from other neural networks by their superior performance with image, speech or audio signal inputs [3].

The key purpose of this work is to classify the Bangla newspaper articles in different categorizes that their consumer are illustrated to full-fill their needs in the best possible ways. The major contribution of this paper are summarized below:

1. In our work, we classify the large Bangla newspaper article dataset which consist of 400k label articles with 25 categorizes in json format.
2. Deep learning approaches for various character have been used for classifying text.
3. In this work we perform to discover different types of statistical analysis.
4. We validate the proposed model's efficiency comparing with state of art and various other existing model.

The rest of the paper is structured as follows section 2 discusses related work on text classification. The methodology are described in section 3 accounts. Experimental findings and statistical analysis are described in section 4. This article is concluded in the last section 5

Related Work

This section represent related work work our proposed model. Many great contributors had already placed a significant role in the field of Bangla article classification. In paper [4] represent sentiment analysis using LSTM and CNN on IMDB comments. This paper proposed a model where a large number of CNN-LSTM layers are combined for analysis. In recent past many contributors have performed multi-label classification in Bengali sentences [5, 6]. We have found one related research paper [7] where authors had performed almost 10,000 Bengali sentence with total of eight labels collected from crime type news articles for multi-label classification. The paper [8] represent word2vec embedding with SVM is utilized for categorizing English documents. Compared to other language, a very few works have been addressed done for Bangla text classification. In our proposed model we use Long Short Term Memory (LSTM) and Convolutional Neural Network (CNN). In paper [9] describes deep learning models such as CNN which is very connected to our proposed model. Moreover, various deep learning algorithm such CNN and LSTM have also been extract valuable features from the unstructured textual data in order to categorize the documents. Naïve Bayes [10] and KNN [11] have been employed in the state-of-the-art works. The paper [12] describes us that a unique N-gram technique for Bengali along with Marathi and Hindi language have been proposed where the highest map value they have found for Bengali text in 47.19%.

Another research article [13] has discovered the usage of N-gram for checking the correctness of Bengali words. In [14] authors have described a model that detects Bengali stop words. This experiment shows an impressive result in precision (100%) where their accuracy is above average (75% at top). We have also studied paper [15] which refers that authors have presented a practical Bengali parts of speech tagger. Last but least we've also studied other Bengali related research articles [16] of different domain. These paper are related to sentiment analysis of Bengali text which gives us a brief insight into the Bengali text processing technique. In paper [17], author represent multi-label Bangla article classification using ML-KNN algorithm. In paper [18] author represent novel deep learning techniques.

Proposed Methodology

1. Data set Extraction

This segment comprises Bangla news taken from Kaggle dataset [19], it contains newspaper articles from different Bengali newspaper. Articles have been already categorized into different classes like international, sports, entertainment, state with the labels attached. For our experiment we consider nine categorizes Bangladesh, Opinion, Lifestyle, International, Education, Economy, Technology, Sports, Entertainment. The programming was completely implemented using Colab with Pandas Library, a versatile Python language development environment with advanced editing, checking and numerical computation environment.

Table 1: Distribution of the Bangla Articles of the Dataset.

Categories	Articles Count
Bangladesh	232504
Opinion	15699
Life-style	10852
International	30856
Education	9721
Economy	17245
Technology	12116
Sports	49012
Entertainment	30466

After loading dataset in json format, we implement the below operations on it.

2. Data Preprocessing

For data preprocessing, we have followed some steps which will be described below. We divide it into two parts, first part have been performed for data cleaning and last part have been used to prepare for model training.

- Remove punctuation and special characters: After loading our dataset we have removed the punctutation like the comma, semicolon other unicode special character. We have also discared Bengali special characters like dari, date and URL.
- Discard stop words: After removing punctuation and special character, we have also discared stop words from our articles.
- Onehotencoding: Onehotencoding is the most important for data preprocessing step. Onehotencoding is the process where it finds all unique labels from collected articles. All the articles will be considered as a one dimension vector.

After complete these steps we divide the input into two phase.

Phase-I: Long Short Term Memory (LSTM) Feature Extraction.

Phase-II: Convolutional Neural Network (CNN) Features Extraction

3. Long Short Term Memory (LSTM) Feature Extraction

Phase-I is used for classification of Bangla newspaper Articles. This segment is implemented by Long Short Term Memory.

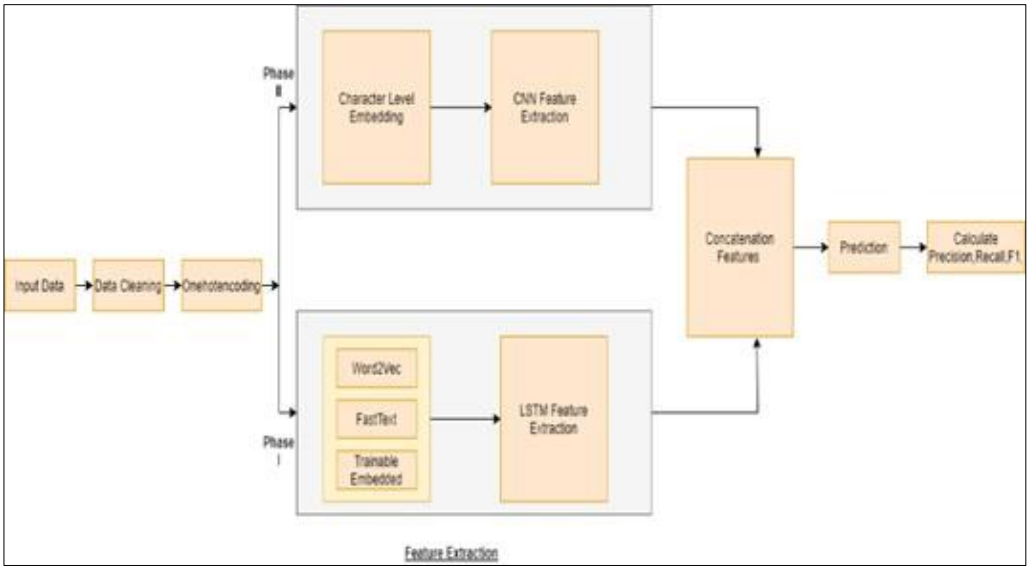

Fig 1: Proposed model of the System

Forget Gate: This gate is useful for the information that no longer useful in the cell state. Two input x_t (input at the particular time) and h_t-1(previous cell output) are fed to the gate and multiplied with weight metrics followed by addition of bias. The result is passed through an activation function which gives a binary output.

Input Gate: The information which is useful of the cell state is done by input gate. First the information is regulated using the sigmoid and filter the values to be remembered similar to the forgate gate using inputs h_t-1 and x_t. Then a vector is created using tanh function that gives output from -1 to 1.

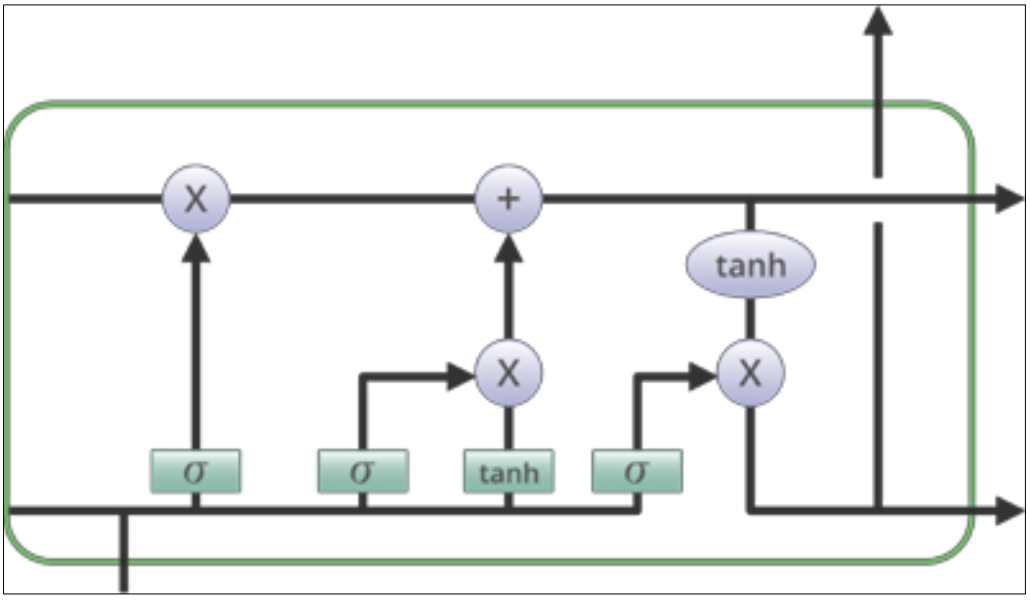

Fig 2: Structure of LSTM [20]

Output Gate: The task of extracting useful information from the current cell state to be presented as an output is done by output gate. First, a vector is generated by applying tanh function on the cell. Then, the information is regulated using the sigmoid function and filter the values to be remembered using inputs h_t-1 and x_t. At last, the values of the vector and the regulated values are multiplied to be sent as an output and input to the next cell.

4. Convolutional Neural Network (CNN) Feature Extraction

Phase-II is intended to contribute to the classification of Bangla newspaper articles. The dataset passes through the preprocessing steps in this segment. Although articles are translated to trainable embedded

numeric vectors, embedding is generated on the basis of all specific characters in the background. Then we apply CNN.

- Character Level Embedded Features: After preprocessing data, articles have been translated to the current vector character level embedded word representation approaches. We used the character level Embed layer to make traniable embedded functionality. We use here 128 dimension embedding layer.
- CNN: In the last stage of phase-II, the trainable embedded features are extracted using CNN. Convolutional Artifical Neural Network is one of the most multi-layer neural network feed-forward model [21]. Figure 3 illustrates the relationship of the layers of the CNN model with the the text as input for evaluating.

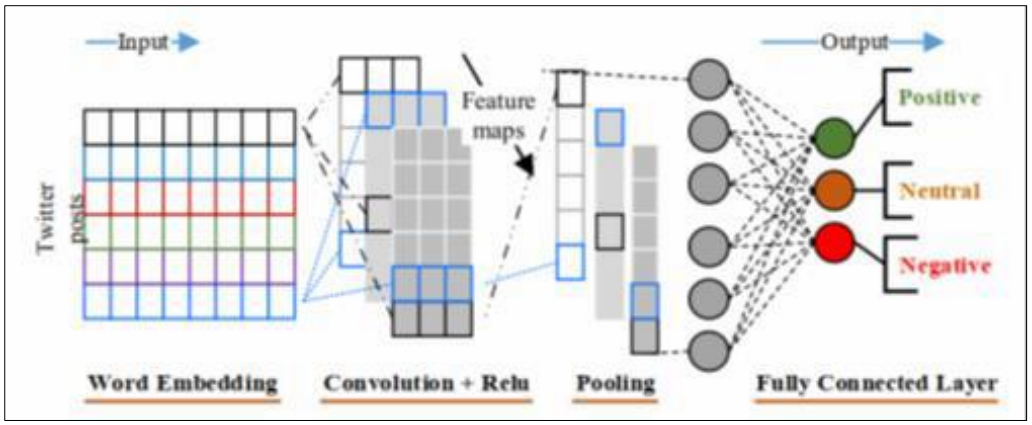

Fig 3: Basic Architecture of CNN [22]

We did a quick experiment, based on the paper by Yoon Kim [9], implementing the 4 Conv Nets models he used to perform sentence classification.
CNN-rand: all words are randomly initialized and then modified during training
CNN-static: pre-trained vectors with all the words— including the unknown ones that are randomly initialized—kept static and only the other parameters of the model are learned
CNN-non-static: same as CNN-static but word vectors are fine-tuned
CNN-multichannel: model with two sets of word vectors. Each set of vectors is treated as a channel and each filter is applied

5. Concatenation Features
At the last stage of the proposed model, phase-I and II features were merged. It allows us to make predictions about all the labeled info. In each of our deep learning classifiers, we use the drop-out layer to avoid a model from over-fitting, as we realize in deep learning module testing, there would be a problem of over-fitting.

Result and Discussion
This section summarizes the experimental findings and performance interpretation of proposed-solution.

1. Performance Metrics
All observations are obtained from uncertainty matrix will be matched with the classification findings obtained in associated tests from the classification to illustrate accuracy. In our experiment we calculate accuracy, precision and F1 are obtained from the confusion matrix.
Accuracy can be measured using the equation

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision refers the approximation of the class labels for each class. Precision can be measured using equation 2.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall value is the weighted average of the right points described correctly in any class. The equation defines this value.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1 is specified in equation 4 and F1 value is close to 1 for a good measure.

$$\text{F1 score} = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

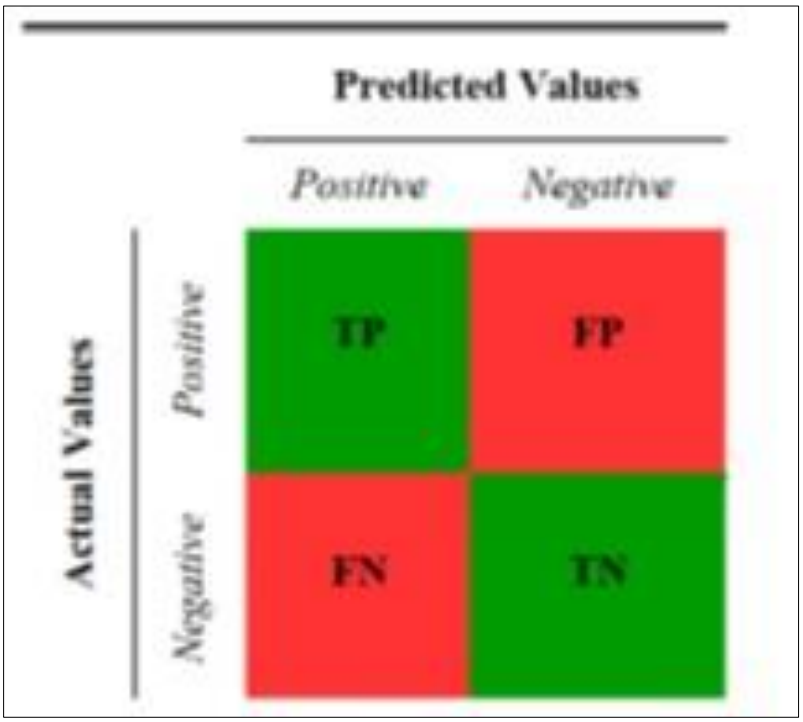

Fig 5: Confusion Matrix [22]

2. Results and Comparison

Many libraries and resources for constructing deep learning model are now open. In this paper we use the accuracy value as the critical success measure to equate our findings. In the accuracy calculation, we consider precision, recall and F1 score to determine the classifiers overall accuracy. Table II depicts the results of our proposed model for classification of Bangla newspaper articles. Again, tne overall accuracy of our model is 95%.

Table 2: Obtained test results of the proposed model for Classification

Category	Predicted Value
Bangladesh	99.697709%
Economy	0.298590%
Education	0.003093%
Entertainment	0.000161%
International	0.000049%
Life-style	0.000065%
Opinion	0.000237%
Sports	0.000024%
Technology	0.000063%

In experiment evaluation we have used 10 fold cross validation. We have trained our model six supervised learning classification techniques such as Decision Tree, Random Forest, SVM, Gaussian Naïve Bayes, Logistic Regression and KNN. We have measured the performance of each supervised learning model using Word2Vec features.

To measure the performance of the prediction model counted Precision, Recall and F1 score as evaluation metrics.

Table 3: Perfomance comparison of different supervised learning model.

Classifiers/Model	Precision	Recall	F1
Decision Tree	69%	69.2%	64.5%
Random Forest	81%	81.6%	81.9%
SVM	78.2%	82.4%	82.8%
Gaussian Naïve Bayes	63.3%	63.7%	63.6%
Logistic Regression	85.6%	83.5%	81.9%
KNN	58%	58.2%	58.3%
Proposed Model	89.51%	92.43%	92.57%

In the experiment evaluation result using proposed model which is depicted in Table 3. We can easily identify that LSTM with CNN is superior to other supervised models. This is the indication that we can utilize the deep neural network learning model to improve the article classification.
In Table 4. Shows us the performance of categorizing Bangla text classification. In our experiment we categorize the Bangla text into nine classes. To categorize the articles we have used performance metrics to evaluate their class.

Table 4: Performance of our proposed model on Whole Dataset.

Category	Precision	Recall	F1
Bangladesh	77%	81%	79%
Opinion	94%	75%	81%
Life-style	93%	72%	78%
International	95%	85%	87%
Education	82%	74%	89%
Economy	88%	77%	72%
Technology	78%	83%	74%
Sports	85%	76%	89%
Entertainment	77%	88%	76%

In Table 5, We compared our experiment performance of our model with other state-of the art works. We apply Long Short Term Memory and Convolutional Neural Network with Word2Vec features from our trained model to perform the comparison, as these models showed better performance in the experiment evaluation. We compare our model with TF-IDF based SVM (Support Vector Machine) Bangla text classification model [6]. This work utilize a dataset of 1000 web documents of five classes. We also compared our model with paper [23], here author used 1960 Bangla web documents of five class. Here author introduced LIBLINEAR model that perform the best and the the average precision is 93%. In paper [24] author represent logistic regression and neural network with both Word2Vec and TF-IDF (3000 features).

Table 5: Performance comparison of different state -of -the art works

Features	**Learning Model**	**Precision**	**Recall**	**F1 Score**
TF-IDF [6]	SVM	0.89	0.89	0.89
TF-IDF [23]	LIBLINEAR	0.93	-	-
Word2vec [24]	Logic Regression	0.95	0.95	0.95
	Neural Network	0.96	0.96	0.96
TF-IDF [24]	Logic Regression	0.94	0.94	0.94
	Neural Network	0.96	0.96	0.96
Proposed Model (word2vec)	LSTM CNN	0.89	0.92	0.92

Conclusion

As we know that multi-label Bangla text classification is very hard job in natural language processing because of it's less research domain. In previous work they categorize the Bangla text classification into five class and very small Bangla web documents in CSV format. But in our work we have categorized the articles into nine classes with a huge size of Bangla newspaper articles dataset which is in JSON format. We have used hybrid model which shows better performance than other supervised learning model techniques. In this work, we curated the largest Bangla newspaper article dataset and performed extensive experiment to compare the performance of different supervised learning model. We expect that this work will help the Bangla natural language research community to further extend the article classification task.

References

1. Brownlee J. 'A Gentle Introduction to Long Short-Term Memory Networks by the Experts', *Machine Learning Mastery*, May 23, 2017. https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/ (accessed Jun. 12, 2021).
2. Van Houdt G, Mosquera C, Nápoles G. 'A review on the long short-term memory model', *Artif. Intell. Rev.*,2020:53(8)5929–5955, Dec. 2020, doi: 10.1007/s10462-020-09838-1.
3. 'What are Convolutional Neural Networks?', Jan. 06, 2021. https://www.ibm.com/cloud/learn/convolutional-neural-networks (accessed Jun. 20, 2021).
4. A. Yenter and A. Verma, 'Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis', in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, Oct, 2017, 540–546. doi: 10.1109/UEMCON.2017.8249013.
5. Dhar A, Mukherjee H, Sekhar Dash N, Roy K. 'Performance of Classifiers in Bangla Text Categorization', in *2018* International Conference on Innovations in Science, Engineering and Technology *(ICISET)*, Oct, 2018, 168–173. doi: 10.1109/ICISET.2018.8745621.

6. Mandal AK, Sen R. 'Supervised learning Methods for Bangla Web Document Categorization', *ArXiv14102045 Cs*, Oct. 2014, Accessed: Jun. 12, 2021. [Online]. Available: http://arxiv.org/abs/1410.2045
7. Md Hasan N, Bhowmik S, Md Rahaman M. 'Multi-label sentence classification using Bengali word embedding model', in *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, Dec. 2017, pp. 1–6. doi: 10.1109/EICT.2017.8275207.
8. J. Lilleberg, Y. Zhu, and Y. Zhang, 'Support vector machines and Word2vec for text classification with semantic features', in *2015 IEEE 14th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, Jul. 2015, pp. 136–140. doi: 10.1109/ICCI-CC.2015.7259377.
9. Y. Kim, 'Convolutional Neural Networks for Sentence Classification', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1746–1751. doi: 10.3115/v1/D14-1181.
10. 'Feature selection for text classification with Naïve Bayes - ScienceDirect'. https://www.sciencedirect.com/science/article/abs/pii/S0957417408003564 (accessed Jun. 21, 2021).
11. V. Tam, A. Santoso, and R. Setiono, 'A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization', in *Object recognition supported by user interaction for service robots*, Aug,2002:4: 235–238. doi: 10.1109/ICPR.2002.1047440.
12. 'Ngram Tokenization for Indian Language Text Retrieval Paul'. https://slidetodoc.com/ngram-tokenization-for-indian-language-text-retrieval-paul/ (accessed Jun. 21, 2021).
13. Hosain Sumit S, Md. Zakir Hossan, T. Al Muntasir, and T. Sourov, 'Exploring Word Embedding for Bangla Sentiment Analysis', in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, 1–5. doi: 10.1109/ICBSLP.2018.8554443.
14. R. ul Haque, Mehera P, Mridha MF, Md A, Hamid. 'A Complete Bengali Stop Word Detection Mechanism', in 2019 Joint 8th International Conference on Informatics, Electronics Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision Pattern Recognition (icIVPR), 2019, 103–107. doi: 10.1109/ICIEV.2019.8858544.
15. K Sarkar and V. Gayen, 'A practical part-of-speech tagger for Bengali', in 2012 Third International Conference on Emerging Applications of Information Technology, 2012, 36–40. doi: 10.1109/EAIT.2012.6407856.
16. Md H Alam, M.-M. Rahoman, and Md. A. K. Azad, 'Sentiment analysis for Bangla sentences using convolutional neural network', in 2017 20th International Conference of Computer and Information Technology (ICCIT), 2017, 1–6. doi: 10.1109/ICCITECHN.2017.8281840.
17. Akanda W, Uddin A. '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), 2021, 466–471. doi: 10.1109/ICICT4SD50815.2021.9396882.
18. 'A Novel Deep Learning based Sentiment Analysis of Twitter Data for US Airline Service'. https://ieeexplore.ieee.org/document/9396879/ (accessed Jun. 25, 2021).
19. 'Bangla Newspaper Dataset'. https://kaggle.com/furcifer/bangla-newspaper-dataset (accessed Jun. 25, 2021).
20. 'Deep Learning | Introduction to Long Short Term Memory', *Geeksfor Geeks*, Jan. 16, 2019. https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/ (accessed Jun. 21, 2021).
21. Yadav A, Vishwakarma DK. 'Sentiment analysis using deep learning architectures: a review', *Artif. Intell. Rev.*,2020:53:6:4335–4385. Aug. 2020, doi: 10.1007/s10462-019-09794-5.
22. MU Salur, Aydin I. 'A Novel Hybrid Deep Learning Model for Sentiment Classification', *IEEE Access*,2020:8:58080–58093. doi: 10.1109/ACCESS.2020.2982538.
23. Dhar A, Dash NS, K Roy. 'Application of TF-IDF Feature for Categorizing Documents of Online Bangla Web Text Corpus', in *Intelligent Engineering Informatics*, Singapore,2018:51-59. doi: 10.1007/978-981-10-7566-7_6.
24. 'BARD: Bangla Article Classification Using a New Comprehensive Dataset | IEEE Conference Publication | IEEE Xplore'. https://ieeexplore.ieee.org/document/8554382 (accessed Jun. 24, 2021).
25. Sudarsan VS, Govind Kuma. Voice call analytics using natural language processing. Int J Stat Appl Math 2019;4(6):133-136.