I. J. Education and Management Engineering, **2021, 6, 39-48**
Published Online December 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2021.06.05

Bangla News Headline Categorization

Amran Hossain
Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh
E-mail: bsse0917@iit.du.ac.bd

Niraj Chaudhary
Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh
E-mail: bsse0836@iit.du.ac.bd

Zahid Hasan Rifad
Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh
E-mail: bsse0820@iit.du.ac.bd

B M Mainul Hossain
Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh
E-mail: mainul@iit.du.ac.bd

Received: 15 May 2021; Revised: 07 July 2021; Accepted: 10 August 2021; Published: 08 December 2021

Abstract: News categorization from various newspapers is important as readers want to read the news by category. But, the readers face difficulty if the news from different categories is presented without any order. This study aims to determine the category of news from online Bangla newspapers. In this context Bangla news headlines data, along with its categories, were collected from various online newspapers through scrapping. Eight categories of news are considered for this work and the headlines of the news are used for categorization. The input data is modeled by the LSTM and GRU neural networks, and the predicted category is compared with the actual category. For LSTM model, the result gives an accuracy of 82.74% and GRU model, The result gives an accuracy of 87.48%. GRU accuracy is higher than LSTM.

Because, GRU training performance is faster than that of LSTM. In GRU 64 units used and in LSTM 128 units used for this research. For this reason, it also suggests that the GRU model gives better results than that of LSTM.

Index Terms: Categorization, Bangla news headlines, neural networks, LSTM, GRU.

1. Introduction

Many methods help the NLP system to understand text and symbols. Text clarification is the process of categorizing the text into a group of words[1]. Text categorization or classification is a way of assigning documents to one or more predefined categories [2]. It is the task of assigning predefined categories to free-text documents. It can provide theoretical views of document collections and has important application in the real world [3]. It helps the users to look for information faster by searching only in the categories, rather than searching the entire information space. The importance of classifying text becomes even more apparent when the information is too big in terms of volume. There are many works of categorization system of news headlines for another language. But there are a few works for the Bangla newspaper. So, we built a system for news categorization for Bangla newspapers. This study will help to make an automatic system, this study introduces classification methods based on machine learning. In these techniques, classifiers are built (or trained) with a set of training documents. The trained classifiers are then used to assign documents to their suitable categories. Amongst the vast information available on the web, we chose the domain of online news because we observed that the current news websites do not provide efficient search functionality based on specific categories and do not support any kind of visualization to analyze or interpret statistics and trends. The fact that news data is published and referenced frequently makes the problem even more relevant. This motivated us to build a system keeping two types of users in mind, the first user is the newsreader who is interested in browsing news articles based on category and the other is the stakeholder or analyst who is interested in analyzing the statistics to identify past and present patterns in news data. Also, Various news Companies want to categorize the news based on published news in a newspaper.

2. Literature Review

Classification approaches favor researchers dealing with real-time data. Researchers did great adventurous research at that time when technical tools were not much available. Some researchers were successful with machine learning classifiers while some of them got privileges from RNN. Through inspiration, this section considers relevant work which has successful accuracy on classifiers that we have used.

Yang li proposed an SVMCNN approach to classifying short text [4]. They applied some machine learning classifiers CNN, SVM, NB, RNN, LSTM. Finally, they got better results with SVM with CNN (SVMCNN) classifier. They got results of about 90% accuracy with the SVMCNN approach

Word embedding has to prepare the analyzing data, Roger Alan Stein claimed word embedding specialty reduces the systems worst performance [5]. Amin Omidvar through their work they had used clickbait online data from the media then processed with a machine learning classifier and Neural network [6].

Jingjing cai had also claimed CNN mostly spreading the area of classifying the vast amount of data [7]. They had clearly described news text classification, emotion analysis, etc. In the whole paper, they had classified Neural Network and present procedure from preprocessing to model classify and outcome.

Tej Bahadur Shahi another respective researcher who did a prediction for self-acting Nepalese news multi-classification [8]. As well as he finished her research to choose machine learning classifiers and neural networks. Machine learning classifiers such as SVM, Naive Bayes are used with multi-layer connectivity. But there is a little bit of an uncomfortable situation with the neural network. During the process, Nepali news text classification was successful 74.65% on behalf of SVM including RBF. But the neural network is the second one on the list with 73% accuracy. Nepali news text classification data volume is a total of 4964 with 20 several types of news. All Deep learning models like neural networks are hungry for the large numerical value of data.

Pranshengit Dhar, Md. Zainal Abedin worked on bangla news headline categorization using optimized machine learning principle [9]. They applied some machine learning classifiers like SVM, Na ïve bayes, Adabost. They got about 81% accuracy.

Sheikh Abujar proposed a neural network-based Bangla news multi-classification system with comparative performance [10]. They prepare about 86 thousand news headlines. They applied some machine learning classifiers like SVM, Logistic Regression, NB, Random Forest, Neural Network. They got about 90% accuracy with Neural Network approaches.

Mayy M. Al-Tahrawi worked on Arabic Text Categorization [11]. They used logistic regression for this research. They applied NB, SVM, KNN algorithms. They got precision of 96.49, recall of 91.67 and F1-measure of 94.0171.

Tehseen Zia worked on Urdu Text Categorization [12]. They used five feature selection methods (information gain, gain ratio, Chi statistics, symmetric uncertain and OneR) and six classifiers (naive Bayes, KNN, support vector machine with linear, polynomial and radial basis kernels and decision tree) for this research. They got 96% f_measure by using Linear SVM with information gain and Chi square as feature selection methods.

Bjorn Gamback worked on hate speech text classification [13]. He admires a convolution neural network. Aimed with CNN they achieved 86.68%. They apply another approach of word embedding which slightly pushes up this value by 7.3% accepted with softmax function, max pooling. Even then values are automatically increased.

3. Methodology

Following procedure is given below:

We collect data from various Bangla newspapers. We used BeautifulSoup python package for scrapping news from website. After collecting data, we remove unnecessary symbols from the datasets and summarize the dataset. We find how many words, documents, unique words per class etc. in this section. Then we find length frequency distribution from the clean datasets. Then prepare the datasets for the model. We used ninety percent data for training and ten percent data for testing. Then labelling the data with encoded sequence. I trained the model with 10 epochs and batch_size 64. Thus, data are prepared for our model. Two deep learning algorithms are used to predict news headlines and compare the results i.e., Long short-term memory (LSTM) and Gated recurrent unit (GRU). We found accuracy, precision, recall, F1_score from these models. After this, the result will be compared.

3.1 LSTM

Long Short-Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They work tremendously well on a large variety of problems and are now widely used. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods is practically their default behavior, not something they struggle to learn! All recurrent neural networks have the form of a chain of repeating modules of neural networks. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer [14].

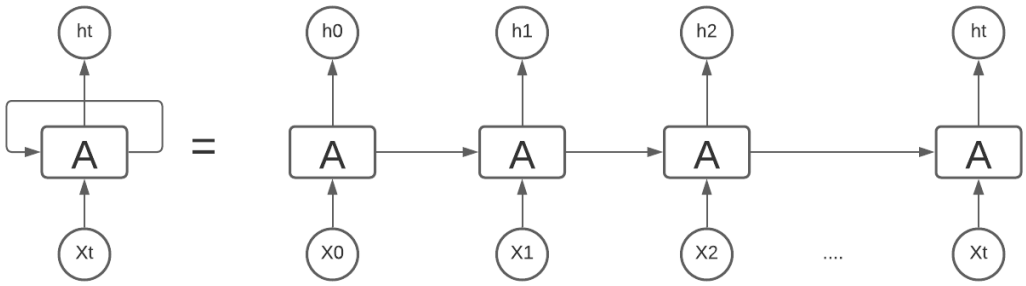


Fig.1. LSTM Architecture (adopted from [15])

The used model architecture consists of an embedding layer (input length =24, embedding dim = 64), LSTM layer (n_unites = 64), two dense layers (n_unites = 24,6), a dropout and a softmax layer.

3.2 GRU

Gated recurrent units (GRUs) are a gating mechanism in recurrent neural networks, introduced in 2014 by Kyunghyun Cho. A GRU has two gates, a reset gate r, and an update gate z. Intuitively, the reset gate determines how to combine the new input with the previous memory, and the update gate defines how much of the previous memory to keep around. If we set the reset to all 1's and update the gate to all 0's we again arrive at our plain RNN model. The basic idea of using a gating mechanism to learn long-term dependencies is the same as in an LSTM.

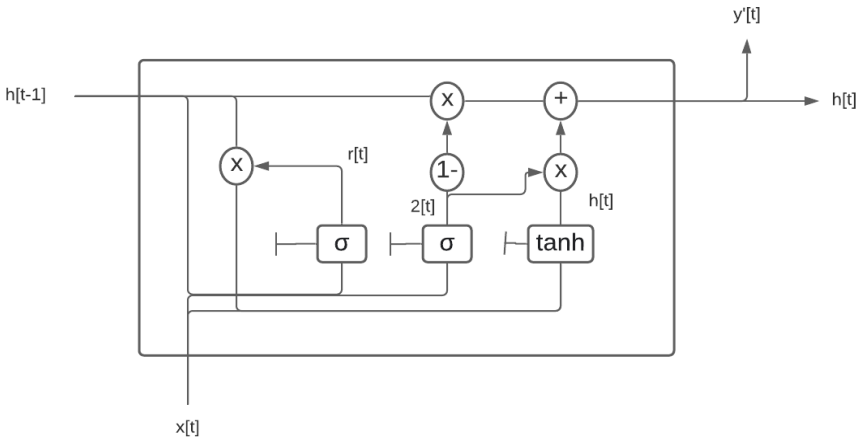


Fig.2. GRU Architecture (adopted from [16])

GRU model has two gates i.e. Update gate and Reset gate to produce the information. It follows some equation to produce relevant information.

Update Gate: The update gate helps the model to determine how much the past information needs to be passed along to the future [12]. This is powerful because the model decides to copy all the information and eliminating the risk of the vanishing gradient problem. For doing this it follows the following equation.

$$z_t = \sigma(W(z)x_t + U(z)h_t - 1) \quad (1)$$

Reset Gate: The reset gate helps the model to determine how much information to forget. The following equation is used to do this.

$$r_t = (W(r)x_t + U(r)h_t - 1) \quad (2)$$

Memory Content: a memory content that will use the reset gate to store the relative information from the past. Two equations are used to store the information from the past and send it to the next time step.

$$h'_t = tanh(W\,x_t + r_t\,o\,U\,h_t - 1) \quad (3)$$

$$h_t = z_t\,o\,h_t + (1 - z)o\,h'_t \quad (4)$$

Here, h'_t is the current memory content and h_t is the final memory content. And *o* means element-wise product. The final h_t is used in the next iteration and continues from (1) to (4).

4. Data

There are many sources to collect data. Newspaper consists of different headlines with different categories. Real-time data is collected from various online newspapers of Bangladesh. Scrapping tools and technology are used for collecting data.

4.1 Data Collection

The data was collected from various Bangla newspapers with scraping. There is more than one lac data in our dataset. We have collected data from various newspapers like Bangladesh pratidin [17], dainik juganttor [18], daily inqilab [19], kalerkantho, and so on. These are the top visited newspapers in Bangladesh. We collected data from these newspapers and it helps this research which categories data are mostly visited to the readers. We used Chrome Web Scrapper and python tools for scraping data from websites. There are three columns in our dataset. These are Headlines, category, and newspaper name. The dataset is publicly available. [1]

The headline distribution of each category represents in the following figure. This dataset is imbalanced.

The following image describes the dataset:

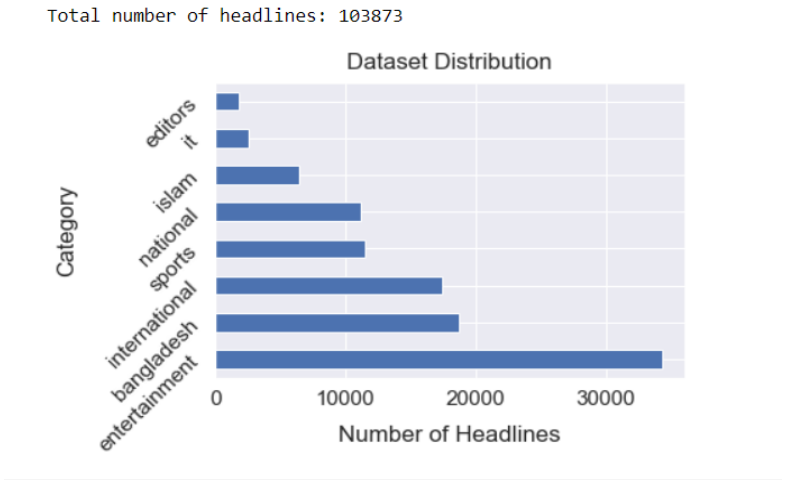


Fig.3. Dataset Description

4.2 Data Cleaning

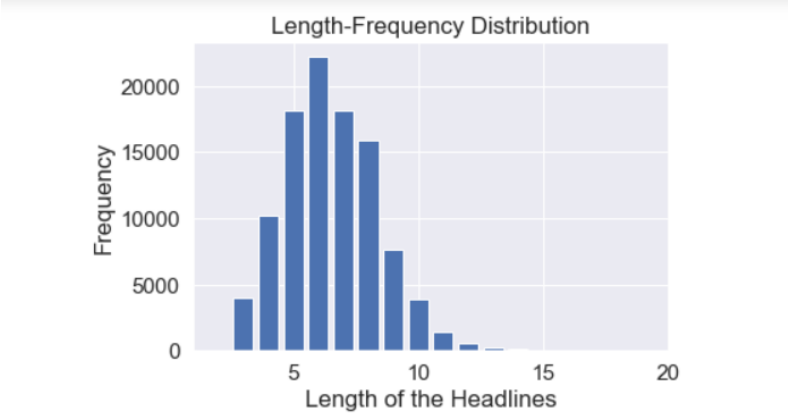
As the headlines are small in length it is not mandatory to remove the stopwords from the headlines [20]. We use regular expressions to remove unnecessary data from our dataset. After cleaning the sample data would look like this.

```
Original:  পদ্মায় ১৪ কেজির রুই, দাম ৩৫ হাজার ৭৫০!
Cleaned: পদ্মায় ১৪ কেজির রুই  দাম ৩৫ হাজার ৭৫০
 Category:-->  bangladesh

Original:  বাংলাদেশের পাখি
Cleaned: বাংলাদেশের পাখি
 Category:-->  editors

Original:  মানবপাচার রোধে বাংলাদেশের জিরো টলারেন্স নীতি গ্রহণ
Cleaned: মানবপাচার রোধে বাংলাদেশের জিরো টলারেন্স নীতি গ্রহণ
 Category:-->  international

Original:  বগুড়ার সংঘর্ষে সাংবাদিকসহ আহত ৭
Cleaned: বগুড়ার সংঘর্ষে সাংবাদিকসহ আহত ৭
 Category:-->  bangladesh

Original:  স্ত্রী-সন্তানসহ করোনায় আক্রান্ত আজিজুল হাকিম
Cleaned: স্ত্রী সন্তানসহ করোনায় আক্রান্ত আজিজুল হাকিম
 Category:-->  entertainment
```

Fig.4. Cleaning Data

[1] https://github.com/amran0917/Bangla-News-Headlines-Categorization/blob/master/headlines2.csv

After cleaning data, we can select the suitable length of headlines we have to use for making every headline into a same length. Figure 5 depicts that maximum, minimum and average length of headlines.




```
Maximum Length of a headline: 19
Minimum Length of a headline: 3
Average Length of a headline: 7.0
```

Fig.5. Length Frequency Distribution

Also, each category has many words. We select unique and similar words from each category. This is called Data Statistics which is depict in Figure 6:

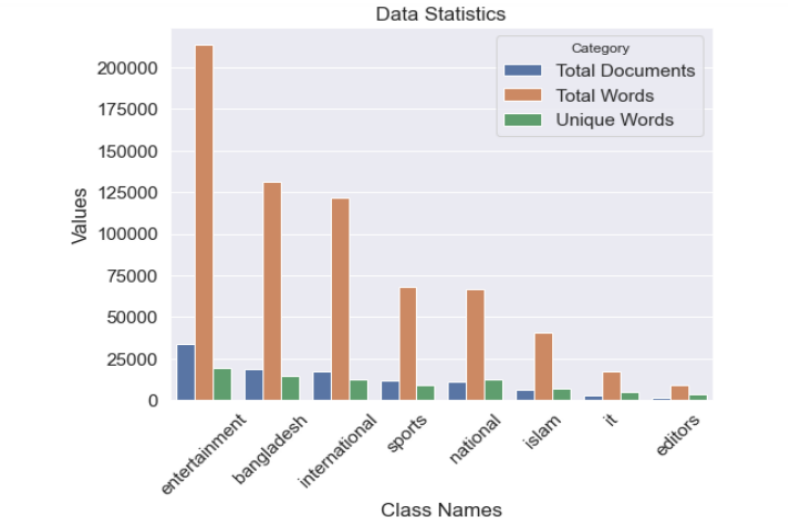


Fig.6. Data Statistics

4.3 Data Preparation and Model Building

The text data are represented by the encoded sequence where the sequences are the vector of an index number that contains words in each headline. The categories are also encoded into numeric values. After preparing the headlines.

```
====== Encoded Sequences ======

শেরপুরে কর্মহীন মানুষের মাঝে খাবার বিতরণ
 [43, 36, 110, 108, 17, 16, 7, 37]

====== Paded Sequences ======
শেরপুরে কর্মহীন মানুষের মাঝে খাবার বিতরণ
[ 43  36 110 108  17  16   7  37   0   0   0   0   0   0   0   0   0   0
  0   0   0]
```

Fig.7. Data Encoding

After padded sequence we add label into each category.

```
                    ===== Label Encoding =====
Class Names:--> ['bangladesh' 'editors' 'entertainment' 'international' 'islam' 'it'
 'national' 'sports']
bangladesh   0

entertainment   2

international   3

bangladesh   0

entertainment   2

entertainment   2

entertainment   2

entertainment   2

entertainment   2
```

Fig.8. After preparing the data

In data processing we collect data from various newspapers with scrapping. Then clean the data using regular expression and others python library. Then cleaned data summarize into category based. Finally prepared date for the model.

Total Data processing is given below:

Fig.9. Data Processing Tasks (step by step)

As result, the analysis dataset is divided into three parts as Test, Train, and Validation. This data distribution is given below

Dataset Distribution:

Set Name	Size
Full	102626
Training	73890
Test	10263
Validation	18473

Fig.10. Dataset Distribution (Test, Train. Validation)

5. Result Analysis and Discussion

We have used two models for predicting news headlines such as LSTM and GRU. We found different results from these two different models. Table 1 discuss about the accuracy of models. GRU Model gives higher accuracy than LSTM Model. For LSTM Model we used 128 units in the contrary GRU model we used 64 units and dropout. For this reason, GRU gives better result. Both are used Bidirectional model and softmax activation function. The better result gives how many data are strictly classified. In GRU Model we accurately classified the categories rather than the LSTM Model.

Table 1. Result analysis

Model	Accuracy
GRU	87.48%
LSTM	82.74%

GRU model gives a better result than the LSTM model. Because GRU is the updated version of LSTM and it is more efficient than LSTM. Also analyzing the performance measure confusion matrix is a best approach. Row determined the actual class and column is predicted class. accuracy, precision, recall, and f1-score are used for measurement. In this study we considered it for measuring performance. These features used in both models in our study.

Accuracy can be measure in the way of equation (5)

$$\text{Accuracy} = (\text{TP} + \text{TN}) \Big/ (\text{TP} + \text{FP} + \text{FN} + \text{TN}) \tag{5}$$

Precision is number of positive results divided by number of all results. The formula of precision in equation (6)

$$\text{Precision, P} = \text{TP} \Big/ (\text{TP} + \text{FP}) \tag{6}$$

Recall measures the number of positive class divided by all positive sample in the dataset. Formula shown in equation (7).

$$\text{Recall, R} = \text{TP} \Big/ (\text{TP} + \text{FN}) \tag{7}$$

F measure is the combination of Precision and Recall. Formula shown in equation (8)

$$\text{F} = 2 * ((\text{Precision} * \text{Recall}) \Big/ (\text{Precision} + \text{Recall})) \tag{8}$$

5.1 LSTM Model

In this simple model, we have got 82.74% validation accuracy for such a multiclass imbalanced dataset. Besides Confusion Matrix and other evaluation measures have been taken to determine the effectiveness of the developed model. From the confusion matrix, it is observed that the maximum number of misclassified headlines is in the category of national, international, and editors and it makes sense because these categories headlines are kind of similar in words.

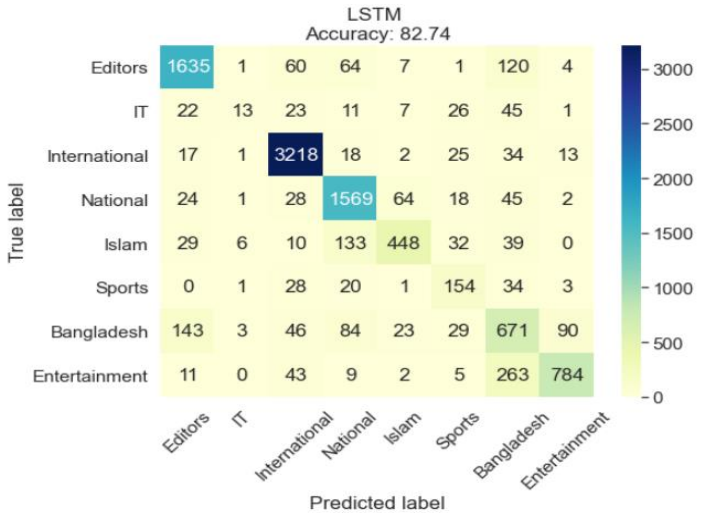


Fig.11. LSTM Model accuracy

The accuracy, precision, recall, and f1-score result also demonstrate this issue. In Fig 12. IT category data's precision, recall, f1-score is lower than others. In the contrary, international category is higher than others.

	precision	recall	f1-score	support
Editors	86.92	86.42	86.67	1892.000000
IT	50.00	8.78	14.94	148.000000
International	93.11	96.69	94.87	3328.000000
National	82.23	89.61	85.76	1751.000000
Islam	80.87	64.28	71.62	697.000000
Sports	53.10	63.90	58.00	241.000000
Bangladesh	53.64	61.62	57.35	1089.000000
Entertainment	87.40	70.19	77.86	1117.000000
accuracy	82.74	82.74	82.74	0.827438
macro avg	73.41	67.69	68.38	10263.000000
weighted avg	82.91	82.74	82.37	10263.000000

Fig.12. LSTM Model (precision, recall, f1-score)

5.2 GRU Model

In this model, the accuracy is about 87.48% which is better than LSTM Model. It is a multiclass imbalanced dataset.

Besides Confusion Matrix and other evaluation measures have been taken to determine the effectiveness of the developed model. From the confusion matrix, we can see the outlook of the result.

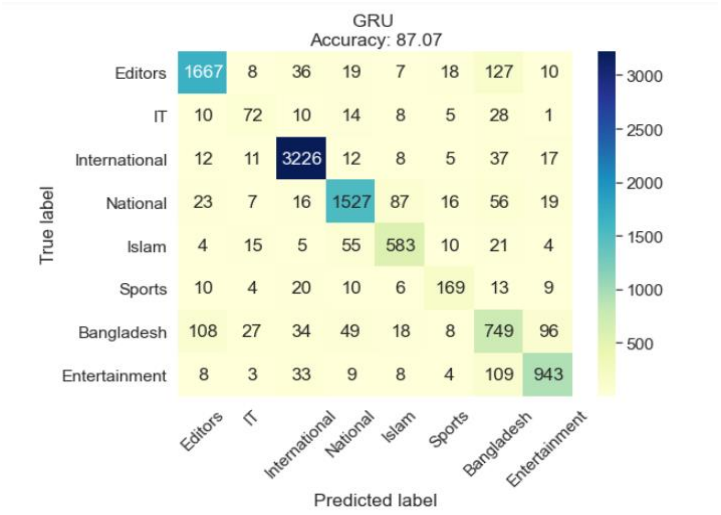


Fig.13. GRU Model Accuracy

The accuracy, precision, recall, and f1-score result also demonstrate this issue. In Fig 14. IT category data's precision, recall, f1-score is lower than others. In the contrary, international category is higher than others.

	precision	recall	f1-score	support
Editors	90.13	89.32	89.73	1892.000000
IT	63.44	39.86	48.96	148.000000
International	94.47	97.03	95.73	3328.000000
National	89.94	87.78	88.84	1751.000000
Islam	82.80	82.21	82.51	697.000000
Sports	75.00	73.44	74.21	241.000000
Bangladesh	66.33	72.18	69.13	1089.000000
Entertainment	87.87	82.99	85.36	1117.000000
accuracy	87.48	87.48	87.48	0.874793
macro avg	81.25	78.10	79.31	10263.000000
weighted avg	87.50	87.48	87.42	10263.000000

Fig.14. GRU Model (precision, recall, f1-score)

5.3 Result Comparison

There are some works in this same field. They apply different methodology like SVM, NB, NN, LR, LSTM, and GRU. They found different results with different methodology. The results are different among the works. We have given

The difference among other works with us which is given below:

Table 2. Result comparison

Model	Accuracy
Our Model	87.48%
IEEE 20115996 [21]	85.14%
IEEE 49239 [22]	90%
Eftekhar Hossain [23]	84%

In IEEE 20115996 they apply LSTM and 8 news categories. Their dataset is less than 1 lac. But we used GRU and 8 categories and our dataset is more than 1 lac. Our model performs better than their model.

IEEE 49239 applies Neural Network (type not defined) and other models like SVM, NB, etc. they got better performance in NN.

Eftekhar Hossain also did the same work with GRU Model and gets 84% accuracy which is less than our model.

6. Conclusion

This paper has derived a machine learning-based model for News headlines Categorization for Bengali newspaper. Most of the studies in the literature consider another linguistic newspaper. GRU is the strongest algorithm for finding a good model for this categorization procedure. The findings from the categorizations are mostly consistent with the literature. As we used two algorithms for this classification, we can differ the result from one model to another. We have taken eight categories for news categorization. The results do not depend on categories. More data, balanced and dissimilar data give a more accurate result for this procedure. Various news Companies want to categorize the news based on published news in the newspaper. So, they may get their results as they want. Overall findings, Further research is required. This dataset is small. So more than dataset will give better result from us. Also changing in model characteristic given different results. For changing epochs result should be changed. Also, without using activation function cause effect in the models.

There are plenty of machine learning models. Using different models give different result.

References

[1] Meparlad, Understanding Text Classification in NLP with Movie Review Example, AnalyticsVidhya, (2020).
[2] Md. Mahmudul hasan shahin, Tanvir Ahmed, Shahriyar Hasan Piyal, Classification of Bangla news articles using bidirectional long short-term memory, (2020).
[3] Yiming Yang and Thorsten Joachims (2008) Text categorization. Scholarpedia, 3(5):4242.
[4] Yang li, Short Text Classification With Convolutional Neural Networks Based Method, (2018).
[5] J. Roger Alan Stenin, Patricia A. Jaques, An Analysis of hierarchical text classification using word embedding (2018).
[6] A. Amin Omidvur, Hui Jiang, Using Neural Network for Identifying Clickbaits in online news media, Communications in computerand information science, 220-232, (2019).
[7] W. J. Jingjing Cai, Jianping Li, Deep Learning model used text classification (2018).
[8] A. Tej Bahadur Shahi, Nepali News Classification using Naïve Bayes, Support Vector Machine and Neural Networks (2018).
[9] Pranshengit Dhar, Md. Zainal Abedin, Bangla News Headline Categorization Using Optimized Machine Learning Principle, (2021).
[10] Sharun Akter Khushbu, Abu Kaiser Mohammad Masum, Neural Network Based Bangali New Headline Multi Classification System: Selection of Features Describes Comparative Performance, (2020).
[11] Mayy M. Al-Tahrawi, Arabic Text Categorization Using Logistic Regression, (2015).
[12] Tehseen Zia, Qaiser Abbas,Muhammad Pervez Akhtar,"Evaluation of Feature Selection Approaches for Urdu Text Categorization", (2015).
[13] Bjorn Gamback, Utpal Kumar Sikdar, “Using Convolutional Neural Networks to Classify Hate-Speech”, (2017).
[14] Oriol Vinyals, Understanding LSTM Networks, Colah.Github, https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (2015).
[15] Simeon Kostadinov, Understanding GRU Networks, Towardsdatascience, https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be (2017).
[16] Bangladesh protidin, https://www.bd-protidin.com (2021).
[17] Doinik Jugantor, https://www.jugantor.com (2021).
[18] Daily Inqilab, https://www.dailyinqilab.com (2021).
[19] Omar Elgabry, The Ultimate Guide to Data Cleaning, Towardsdatascience, https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4 (2019)

[20] J. Shazia Usmani, News Headlines Categorization Scheme for Unlabeled Data, (2020).
[21] Sharun Akter Khushbu, Neural Network Based Bengali News Headline Multi Classification System: Selection of Features describes Comparative Performance, (2020).
[22] Rick Anderson, RNN Talking about Gated Recurrent Unit, https://technopremium.com/blog/rnn-talking-about-gated-recurrent-unit/, (2019).
[23] Eftekhar Hossain, Bangla News Headline Categorization using Gated Recurrent Unit (GRU), Github, (2020).
[24] F Ciravegna, L Gilardoni, A Lavelli, S Mazza, W J Black, M Ferraro, et al, "Flexible text classification for financial applications: the FACILE system," in ECAI, 2000, pp 696-700.
[25] M Taboada, J Brooke, M Tofiloski, K Voll, and M Stede, "Lexiconbased methods for sentiment analysis," Computational linguistics, vol 37, no 2, pp 267–307, 2011 Chen, S Y, & Hsieh, J W Boosted Road sign detection and recognition In Proc of Intl Conference on Machine Learning and Cybernetics, 2008 pp 3823–3826.
[26] A Khan, B Baharudin, and K Khairullah, "Sentiment classification using sentence-level lexical based semantic orientation of online reviews," Trends in Applied Sciences Research, vol 6, no 10, pp 1141–1157, 2011.
[27] Hawalah, Ahmad 2019 "Semantic Ontology-Based Approach to Enhance Arabic Text Classification "Big Data Cogn Comput 3", no 4:53.

Authors' Profiles

Amran Hossain is an undergrad student of Institute of Information Technology in Software Engineering major at the University of Dhaka, Dhaka, Bangladesh.

Niraj Chaudhary is an undergrad student of Institute of Information Technology in Software Engineering major at the University of Dhaka, Dhaka, Bangladesh.

Zahid Hasan Rifad is an undergrad student of Institute of Information Technology in Software Engineering major at the University of Dhaka, Dhaka, Bangladesh

Dr. B M Mainul Hossain is Associate Professor at Institute of Information technology, University of Dhaka. He received his Ph.D. from the University of Illinois at Chicago. He completed his MSc & BSc degrees from the Department of Computer Science and Engineering, University of Dhaka.

How to cite this paper: Amran Hossain, Niraj Chaudhary, Zahid Hasan Rifad, B M Mainul Hossain, " Bangla News Headline Categorization", International Journal of Education and Management Engineering (IJEME), Vol.11, No.6, pp. 39-48, 2021. DOI: 10.5815/ijeme.2021.06.05