Bangla News Headline Categorization

Amran Hossain*, Niraj Chaudhary*, Zahid Hasan Rifad*, Dr. B M Mainul Hossain*

Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh

Abstract

The modern era is developing in all sectors. Moreover development is necessary to research. Accordingly natural language processing(NLP) is the expanding area of process text. This project. At present year, memorizing the data is very tough due to the rapidly growing volume of data. Newspapers are a great habit to the whole world of all ages. To acquire a variety of knowledge from different segments is entertaining by themselves. According we categorizes news headlines from various Bengali newspapers. In addition, it also lets one classify newspapers based on the category they may correspond to ¬ national, international, sports etc. So, we wanted to do this work for Bengali newspapers. We worked for eight categories from the newspapers and completed the classification of news headlines.We apply LSTM and GRU for predicting news headline and it give upper than 85% accuracy.

Keywords: natural language processing, classification, bangla news

*Corresponding author

Email addresses: bsse0917@iit.du.ac.bd (Amran Hossain), bsse0836@iit.du.ac.bd (Niraj Chaudhary), bsse0820@iit.du.ac.bd (Zahid Hasan Rifad), mainul@iit.du.ac.bd (Dr. B M Mainul Hossain)

headlines, deep learning, LSTM, GRU

1. Introduction

[1] Text categorization or classification, is a way of assigning documents to one or more predefined categories. This helps the users to look for information faster by searching only in the categories they want to, 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 is categorization system of news headlines for another languages. But there is no work for Bengali newspaper. So, we built a system for news categorization for Bengali newspapers. In our project, to make this system automatic, classification methods based on machine learning have been introduced. 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 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 on a frequent basis 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 news reader 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 Company want to categorize the news based on published

news on newspaper.

1.1. Literature Review

Classification approaches do favour when researchers dealing with real time data. they did great adventurous research on that time when technical tools were not much available. Some researchers are successful with machine learning classifiers and some of then got privilege from RNN. By means of inspiration this section we consider relevant work which has successful accuracy on classifiers what we have used. Bangla news headline dataset together with model analysis that our model approach and accuracy have uniqueness.

Yang li et al. [2] proposed a SVMCNN approach to classify short text. He applied some machine learning classifiers CNN, SVM, NB, RNN, LSTM. Finally he got better result with SVM with CNN (SVMCNN) classifier. They got resut about 90% accuracy with SVMCNN approach. On behalf of measure with output we take an overview from there that applying neural networks it must behave as superior.

Word embeddings have a duty to prepare the analysing data, Roger Alan Stein et al. [3] claimed word embeddings speciality reduce the systems worst performance. Amin Omidvar et al. [4] through their work they had used clickbaits online data from the media then processed with machine learning classifier and Neural network. We simply follow their strategy of better understanding because they made up with the highest accuracy among the neural network and gained 98% accuracy. Jingjing cai et al. [5] also had claimed CNN mostly spreading the area of classifying the vast amount of data. They had clearly described news text classification, emotion analy-

sis etc. In the whole paper they classified by Neural Network and present procedure from preprocessing to model classify and outcome.

Tej Bahadur Shahi et al. [6] another respective researcher who did prediction for self-acting nepali news multi classification. As well as he finished her research to choose machine learning classifiers and neural networks. Machine learning classifiers such as SVM, Naive bayes 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 to the list with 73% accuracy. Nepali news text classification data volume is total 4964 with 20 several types of news. All Deep learning models like neural networks are hungry for large numerical value of data.

Sheikh Abujar [7] proposed a neural network based bengali news multi-classification system with comparative performance. They prepare about 86 thousands news headline. They applied some machine learning classifiers like SVM, Logistic Regression, NB, Random Forest, Neural Network. They got about 90% accuracy with Neural Network approaches.

So after reviewing this approaches and model we have decided to classify with neural network like LSTM, GRU.

2. Methodology

Mainly two deep learning algorithms are used to predict news headlines. Long short-term memory(LSTM) and Gated recurrent unit(GRU). After this result will be compared between these algorithms.

2.1. LSTM

[8] 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 of time 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 network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

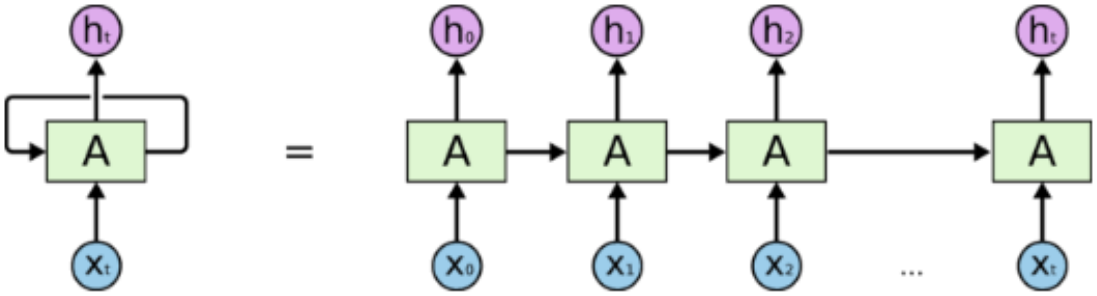


Figure 1: LSTM architecture

The used model architecture consists of a embedding layer (input length =24, embedding dim = 64), LSTM layer ($\mathrm{n}_u nites = 64$), $twodenselayer(n_u nites = 24, 6)$, $adropoutandasoftmaxlayer$.

2.2. GRU

Gated recurrent units (GRUs) are a gating mechanism in recurrent neural networks, introduced in 2014 by Kyunghyun Cho et al [9].A GRU has

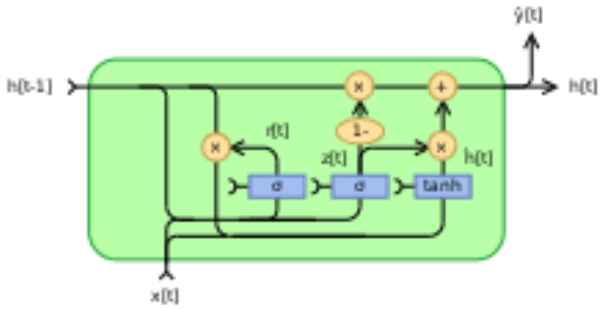


Figure 2: Gated Recurrent Unit

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 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 a LSTM.

2.2.1. Update Gate

We start with calculating the update gate z_t for time step t using the formula:

$$Z_t = \sigma(W^{(z)}x_t + U^{(z)}h_{t-1}) \tag{1}$$

When x_t is plugged into the network unit, it is multiplied by its own weight $W^{(z)}$. The same goes for $4h_{(t-1)}$ which holds the information for the previous t-1 units and is multiplied by its own weight $U^{(z)}$. Both results are added together and a sigmoid activation function is applied to squash the result between 0 and 1.

The update gate helps the model to determine how much of the past in-

formation (from previous time steps) needs to be passed along to the future. That is really powerful because the model can decide to copy all the information from the past and eliminate the risk of vanishing gradient problem.

2.2.2. Reset Gate

Essentially, this gate is used from the model to decide how much of the past information to forget. To calculate it, we use:

$$R_t = \sigma(W^{(r)}x_t + U^{(r)}h_{t-1}) \tag{2}$$

This formula is the same as the one for the update gate. The difference comes in the weights and the gate's usage.

2.2.3. Current Memory Content

A new memory content which will use the reset gate to store the relevant information from the past.

$$h'_t = tanh(Wx_t + r_t * Uh_{t-1}) \tag{3}$$

Multiply the input x_t with a weight W and $h_(t-1)$ with a weight U. the Hadamard (element-wise) product between the reset gate r_t and $Uh_(t-1)$. That will determine what to remove from the previous time steps. *tanh* is the nonlinear activation function.

2.2.4. Final Memory at Current Time Step

the network needs to calculate h_t — vector which holds information for the current unit and passes it down to the network. In order to do that the update gate is needed. It determines what to collect from the current

memory content — h'_t and what from the previous steps — $h_(t-1)$. That is done as follows:

$$h_t = z_t * h_t + (1 - z) * h'_t \tag{4}$$

Element-wise multiplication to the update gate z_t and $h_(t-1)$.Element-wise multiplication to $(1 - z_t)$ and h'_t.

3. Data

Data collected from various news papers. Scrapping tools and technology is used for collecting data.

3.1. Data Collection

The data was collected from various bangla newspaper with scraping. There is more than one lac data in our dataset. We have collected data various newspapers like Bangladesh pratidin [10], dainik juganttor [11], daily inquilab [12], kalerkantho and so on. We used Chrome Web Scrapper and python languages for scrapping data from websites. There are three columns in our dataset. These are Headlines, category and newspaper_name. The headlines distribution of each categories represents in the following figure. This dataset is an imbalanced dataset.

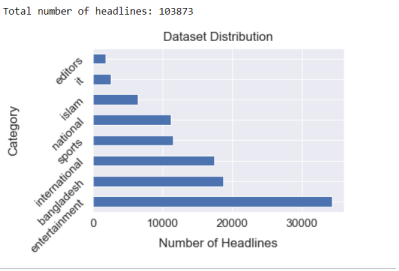
3.2. Data Clean

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

	headline	category	newspaper name
0	মাদারীপুরে নদীতে গোসল করতে গিয়ে শিশুর মৃত্যু	bangladesh	BDProtidin
1	আখাউড়া ইমিগ্রেশন দিয়ে ভারত থেকে আসত ১৪ যাত্রী ...	bangladesh	BDProtidin
2	অবৈধ বিক্রি বন্ধে বালুর স্তূপ মিনামে তুললেন ইউএনও	bangladesh	BDProtidin
3	গরু চুরির অভিযোগে যুবককে নির্যাতন, ইউপি সদস্য ...	bangladesh	BDProtidin
4	মশীগ্রামে খালেদা জিয়ার রোগমুক্তি কামনায় মসজি...	bangladesh	BDProtidin
...	...	...	...
103868	ব্লকচেইন নিয়ে সেমিনার করল বিসিএস	it	jugantor
103869	উবারের দক্ষিণ এশিয়ার প্রেসিডেন্ট হলেন প্রভজিৎ সিং	it	jugantor
103870	এই ঈদে বিরাট হাট বসছে দেশীসরু বিডিডটকমে	it	jugantor
103871	প্রবাসীরাও পশু অর্ডার করতে পারবেন অনলাইন হাটে	it	jugantor
103872	বিনামূল্যে প্রশিক্ষণ পাবেন পাঁচ হাজার জন	it	jugantor

103873 rows × 3 columns

(a) Data Description



(b) Data distribution

Figure 3: Data Collection

```
Original:  পদ্মায় ১৪ কোজির 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
```

Figure 4: Remove unnecessary symbols

3.3. Data Summary

Data summary includes the information about number of documents, words and unique words have in each category class. Also, include the length distribution of the headlines in the dataset.

From this graphical information we can select the suitable length of headlines we have to use for making every headline into a same length.

From this graphical information we can understand document wise word frequency distribution for each headline. It is a data statistics graphical representation.

```
Class Name :  entertainment
Number of Documents:33717
Number of Words:213598
Number of Unique Words:19121
Most Frequent Words:

নতুন      2588
নিয়ে      2288
ও          1802
গান        1381
খান        1074
না         1015
র          1012
নাটক       890
মুক্তি     844
চলচ্চিত্র  838
```

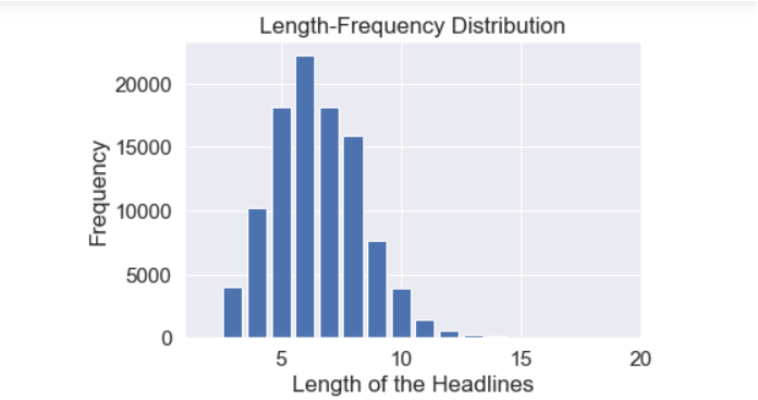
(a) entertainment catgory

```
Class Name :  bangladesh
Number of Documents:18665
Number of Words:131368
Number of Unique Words:14722
Most Frequent Words:

মৃত্যু     1126
নিহত       1041
গ্রেফতার   976
আটক       891
ও          830
২          686
লাশ        671
উদ্ধার     659
থেকে       622
আহত       615
```

(b) Bangladesh Category

Figure 5: Data summary




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

Figure 6: Length Frequency distribution

3.4. Data preparation and model building

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

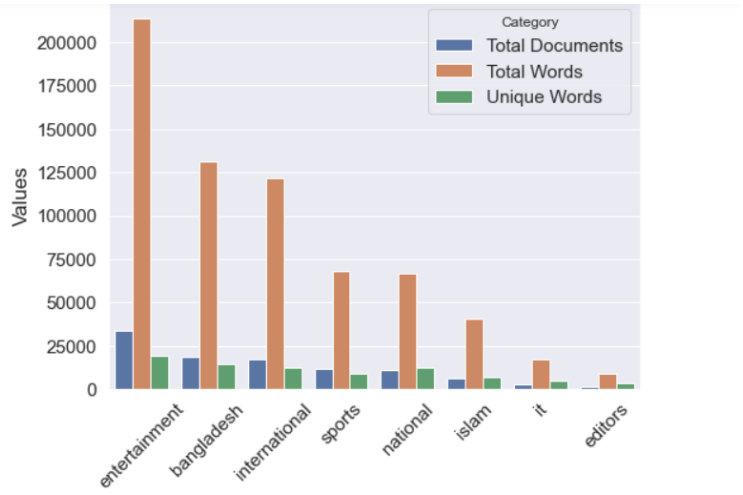


Figure 7: Data Statistics Summery

```
====== 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]
```

Figure 8: Data encoding

And label encoding looks likes the following figure which is given below:

For model evaluation encoded headlines are splitted into train test validation set. The distribution has-

```
                    ===== 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
```

Figure 9: Label encoding

Dataset Distribution:

Set Name	Size
Full	102626
Training	73890
Test	10263
Validation	18473

Figure 10: Test and train dataset

4. Result Analysis and Discussion

Model	Accuracy
GRU	87.48%
LSTM	82.74%

4.1. LSTM Model

In this simple model we have got 82.74% validation accuracy for such a multiclass imbalanced dataset. Besides Confusion Matrix and other evalua-

tion 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 are fall in the category of national, international and editors and it makes senses because these categories headlines are kind of similar in words. The accuracy, precision, recall and f1-score result also demonstrate this issue.

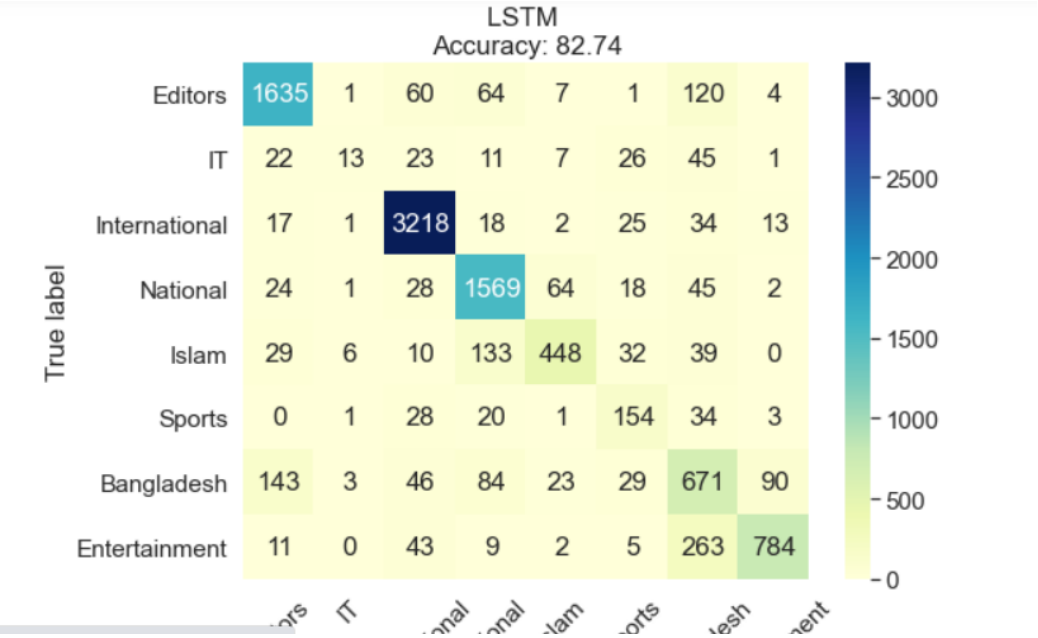


Figure 11: LSTM accuracy

	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

Figure 12: LSTM precision, recall and f1-score

4.2. GRU Model

In this model the accuracy is about 87.48% which is better than LSTM Model. The accuracy, precision, recall and f1-score result also demonstrate this issue.

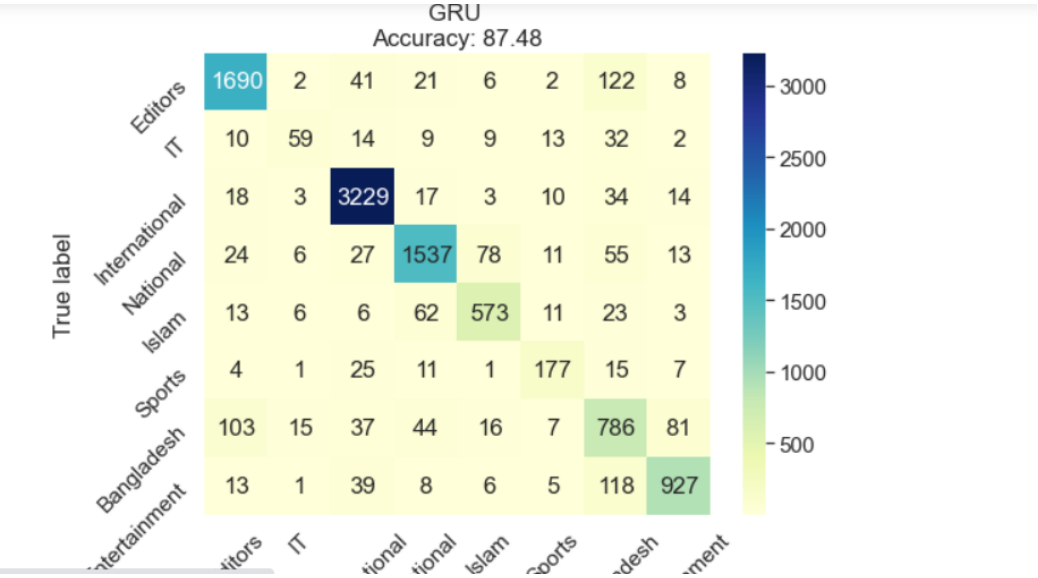


Figure 13: GRU accuracy

	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

Figure 14: GRU precision, recall and f1-score

4.3. Result Comparison

Model	Accuracy
Our Model	87.48%
IEEE 20115996[14]	85.14%
IEEE 49239[15]	90%
eftekhar hossain[16]	84%

5. 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 for another linguistic newspaper. The paper differs it for Bengali newspaper. GRU is the strongest algorithm for finding 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 model. 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 Company want to categorize the news based on published news on newspaper. So, they may get their results as they want.

References

[1] M. Md. Mahmudul Hasan Shahin, Tanvir Ahmmed Shahriar Hasan Piyal, Classification of bangla news articles using bidirectional long short term memory, Tech. rep. (2020).

[2] Yang Li, Short Text Classification With A Convolutional Neural Networks Based Method, https://www.researchgate.net/publication/331701896_Short_Text_Classification_With_A_Convolutional_Neural_Networks_Based_Method (2018).

[3] J. Roger Alan Steina, Patricia A. Jaques, An analysis of hierarchical text classification using word embeddings (2018).

[4] A. A. Amin Omidvar, Hui Jiang, Multi-variate flood damage assessment: a tree-based data-mining approach, Communications in Computer and Information Science (1) (2019) 220–232.

[5] W. J. Jingjing Cai, Jianping Li, Deep Learning Model Used in Text Classification (2018).

[6] A. Tej Bahadur Shahi, Nepali News Classification using Naïve Bayes, Support Vector Machines and Neural Networks (2018).

[7] S. S. Sharun Akter Khushbu, Abu Kaisar Mohammad Masum, Neural Network Based Bengali News Headline Multi Classification System: Selection of Features describes Comparative Performance.

[8] Oriol Vinyals, Understanding LSTM Networks, https://colah.github.io/posts/2015-08-Understanding-LSTMs/ (2015).

[9] Simeon Kostadinov, Understanding GRU Networks, https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be (2017).

[10] Bangladesh Protidin, https://www.bd-pratidin.com/ (2021).

[11] , Doinik Jugantor, https://www.jugantor.com/ (2021).

[12] Daily Inqilab, https://www.dailyinqilab.com (2021).

[13] Omar Elgabry, The Ultimate Guide to Data Cleaning, https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4 (2019).

[14] J. Shazia Usmani, News Headlines Categorization Scheme for Unlabelled Data (2020).

[15] Sharun Akter Khushbu, Neural Network Based Bengali News Headline Multi Classification System: Selection of Features describes Comparative Performance (2020).

[16] Eftekhar Hossain, Bangla News Headlines Categorization Using Gated Recurrent Unit (GRU), https://github.com/eftekhar-hossain/Bangla-News-Headlines-Categorization (2020).