Bangla News Headline Categorization

**Amran Hossain1 and Niraj Chaudhary2 and Zahid Hasan Rifad3 and Dr. B M Mainul Hossain4**

1Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh;

2Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh;

3Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh;

4Institute of Information Technology, University of Dhaka, Dhaka, Bangladesh

E-mail: 1bsse0917@iit.du.ac.bd, 2bsse0836@iit.du.ac.bd, 3bsse0820@iit.du.ac.bd, 4mainul@iit.du.ac.bd

Received: ….. ; Accepted: ….. ; Published: …..

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 Bangla newspapers. In this context 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 the purpose of categorization. The input data is modeled by the LSTM and GRU neural networks, and the predicted category is compared with the actual category. The result gives an accuracy of 85%. It also suggests that the GRU model gives better results than that of LSTM.

Index Terms: natural language processing, classification, Bangla news headlines, deep learning, LSTM, GRU.

##### **1. Introduction**

Text categorization or classification, is a way of assigning documents to one or more predefined categories [1]. It 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 Bangla newspaper. So, we built a system for news categorization for Bangla 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.

##### **2. Literature Review**

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

Yang li proposed a SVMCNN approach to classify short text [2]. They applied some machine learning classifiers CNN, SVM, NB, RNN, LSTM. Finally they got better result with SVM with CNN (SVMCNN) classifier. They got result 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 embedding have a duty to prepare the analyzing data, Roger Alan Stein claimed word embedding specialty reduce the systems worst performance [3]. Amin Omidvar through their work they had used clickbaits online data from the media then processed with machine learning classifier and Neural network [4]. 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 had also claimed CNN mostly spreading the area of classifying the vast amount of data [5]. 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 prediction for self-acting Nepalese news multi classification [6]. 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 proposed a neural network based bangla news multi-classification system with comparative performance [7]. 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.

##### **3. Methodology**

##### 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). 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 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 [8].

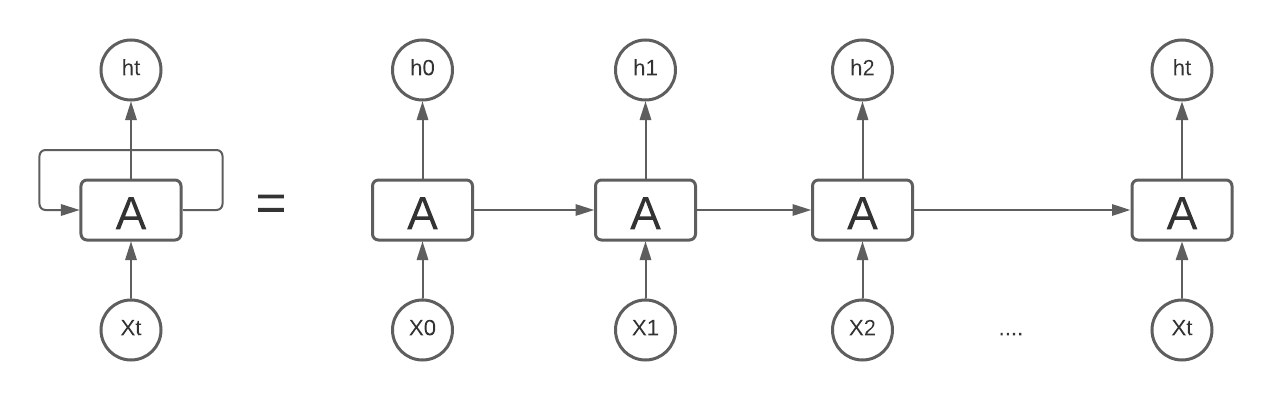


Fig 1: LSTM Architecture (adopted from [8])

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 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.

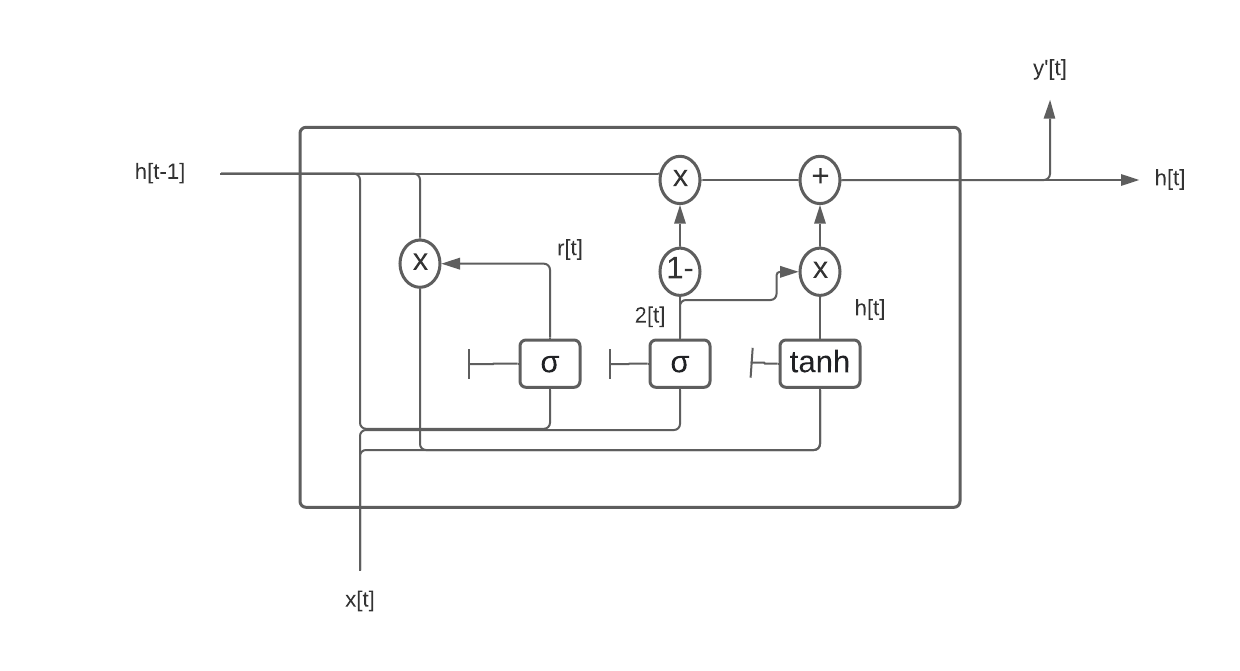


Fig 2: GRU Architecture (adopted from [16])

GRU model has two gate 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 [9]. This is powerful because the model is decide to copy all the information and eliminating the risk of vanishing gradient problem. For doing this it follow the following equation.

zt = (W(z)xt + U(z)ht-1) (1)

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

rt = (W(r)xt + U(r)ht-1) (2)

Memory Content: a memory content which 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 xt + rt o U ht-1) (3)

ht = zt o ht + (1-z)o h′t  (4)

Here, h′t is the current memory content and *ht* is the final memory content. And *o* means element wise product. Final ht is used in the next iteration and continue from (1) to (4).

##### **4. Data**

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

## 4.1 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 from various newspapers like Bangladesh pratidin [10], dainik juganttor [11], daily inquilab [12], kalerkantho and so on. We used Chrome Web Scrapper and python tools 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.

The following image describe the dataset:

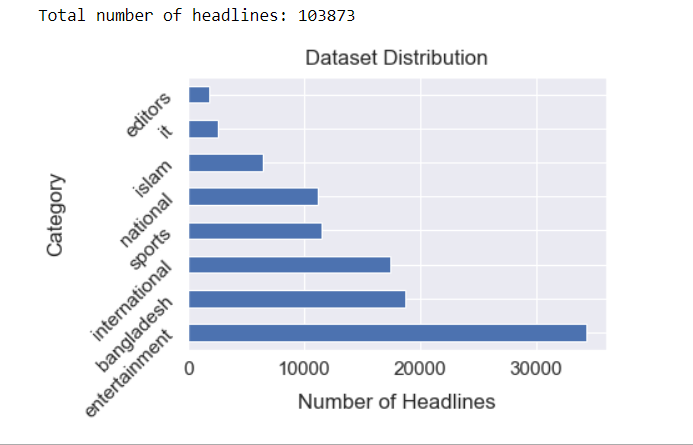


Fig 3: Dataset Description

## 4.2 Data Cleaning

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

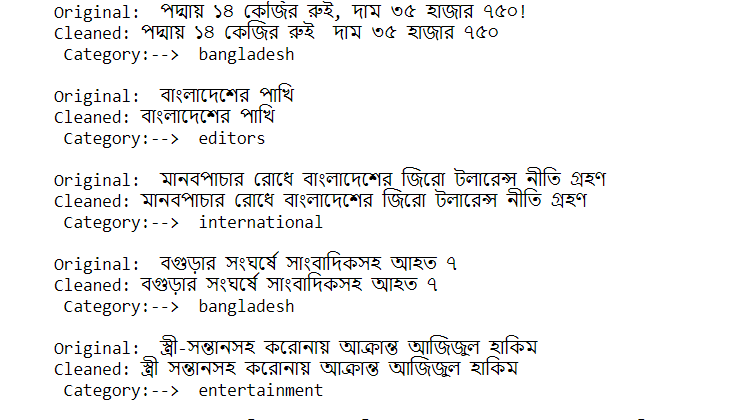


Fig 4: Cleaning Data

## 4.3 Data Preparation and Model Building

The text data are represented by 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.





Fig 5: after preparing the data

Total Data processing is given below:

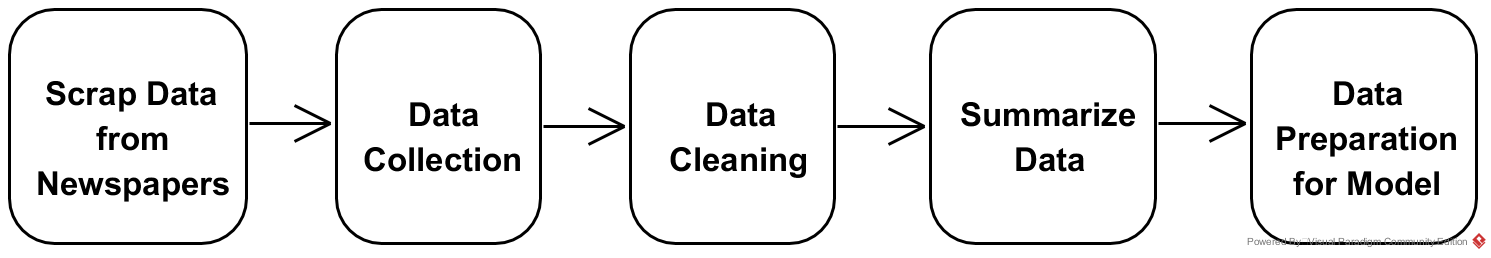


Fig 6: Data Processing Tasks (step by step)

For result analysis dataset is divided in three part such as Test, Train and Validation. This data distribution is given below

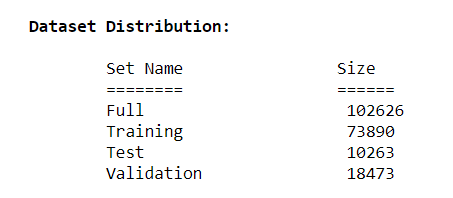


Fig 7: Dataset Distribution (Test, Train. Validation)

##### **5. Result Analysis and Discussion**

We have used two model for predicting news headlines such as LSTM and GRU. We found different result from this two different model.

|  |  |
| --- | --- |
| Model | Accuracy |
| GRU | 87.48% |
| LSTM | 82.74% |

GRU model gives the better result than LSTM model. Because GRU is the update version of LSTM and it is more efficient than LSTM.

## 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 are 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.

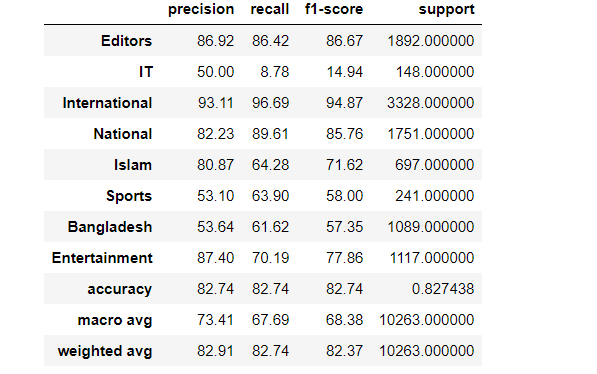


Fig 8: LSTM Model (precision, recall, f1-score)

## 5.2 GRU Model

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

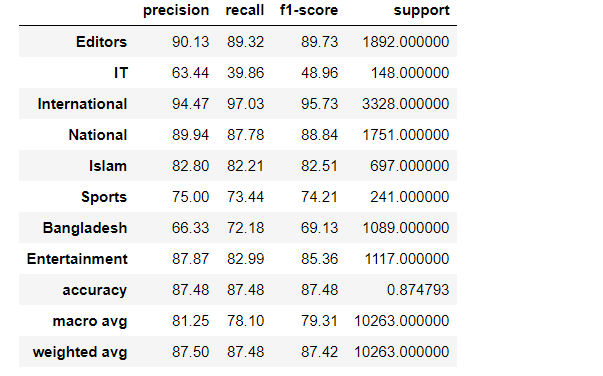


Fig 9: GRU Model (precision, recall, f1-score)

## 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 result with different methodology. Some works with result and comparison with our model are:

|  |  |
| --- | --- |
| Model | Accuracy |
| Our Model | 87.48% |
| IEEE 20115996 [14] | 85.14% |
| IEEE 49239 [15] | 90% |
| Eftekhar Hossain [17] | 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 perform better than their model.

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

Eftekhar Hossain also did the same work with GRU Model and get 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 for another linguistic 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. 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. Md. Mahmudul hasan shahin, Tanvir Ahmed, Shahriyar Hasan Piyal, Classification of Bangla news articles using bidirectional long short term memory, (2020).
2. Yang li, Short Text Classification With a Convolutional Neural Networks Based Method, (2018).
3. J. Roger Alan Stenin, Patricia A. Jaques, An Analysis of hierarchical text classification using word embedding (2018).
4. A. Amin Omidvur, Hui Jiang, Using Neural Network for Identifying Clickbaits in online news media, Communications in computerand information science, 220-232, (2019).
5. W. J. Jingjing Cai, Jianping Li, Deep Learning model used text classification (2018).
6. A. Tej Bahadur Shahi, Nepali News Classification using Naïve Bayes, Support Vector Machine and Neural Networks (2018).
7. Sharun Akter Khushbu, Abu Kaiser Mohammad Masum, Neural Network Based Bangali New Headline Multi Classification System: Selection of Features Describes Comparative Performance, (2020).
8. Oriol Vinyals, Understanding LSTM Networks, Colah.Github, <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> (2015).
9. Simeon Kostadinov, Understanding GRU Networks, Towardsdatascience, <https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be> (2017).
10. Bangladesh protidin, <https://www.bd-protidin.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, Towardsdatascience, <https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4> (2019)
14. J. Shazia Usmani, News Headlines Categorization Scheme for Unlabeled Data, (2020).
15. Sharun Akter Khushbu, Neural Network Based Bengali News Headline Multi Classification System: Selection of Features describes Comparative Performance, (2020).
16. Rick Anderson, RNN Talking about Gated Recurrent Unit, <https://technopremium.com/blog/rnn-talking-about-gated-recurrent-unit/>, (2019).
17. Eftekhar Hossain, Bangla News Headline Categorization using Gated Recurrent Unit (GRU), Github, (2020)

**Authors’ Profiles**

|  |  |
| --- | --- |
| C:\Users\niraj\Downloads\24257.JPG | **Amran Hossain** is an undergrad student of Institute of Information Technology in Software Engineering major in University of Dhaka, Dhaka, Bangladesh. |
| E:\photos\pp photo\IMG_2528=600x600.jpg | **Niraj Chaudhary** is an undergrad student of Institute of Information Technology in Software Engineering major in University of Dhaka, Dhaka, Bangladesh. |
| **C:\Users\niraj\Downloads\my-passport-photo.jpg** | **Zahid Hasan Rifad** is an undergrad student of Institute of Information Technology in Software Engineering major in University of Dhaka, Dhaka, Bangladesh |
| C:\Users\niraj\Downloads\BMH.jpg | **Dr. B M Mainul Hossain** is Associate Professor at Institute of Information technology, University of Dhaka. He received his PhD 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:** Dr. B M Mainul Hossain, Amran Hossain, Niraj Chaudhary, Zahid Hasan Rifad, "Bangla News Headline Categorization", International Journal of Information Technology and Computer Science (IJITCS), Vol.1, No.1, pp.1-7, 2021. DOI: 10.5815/ijitcs.2021.05.12