Ahsanullah University of Science and Technology
Department of Computer Science and Engineering

CSE 4238: Soft Computing Lab

Spring 2021

Project Report

Bangla News Headline Categorization

Lab Section: B1

Submitted by

Ashiqul Islam	170204070
Alam Khan	170204084
Mehedi Hasan	170204096

Submitted To

Mr. H M Zabir Haque	Mr. Nibir Chandra Mandal
Assistant Professor	Lecturer
Department of CSE, AUST	Department of CSE, AUST

Date of Submission: 08 October, 2022

Contents

1 Introduction

Text classification process involves data collection, pre-processing, feature selection, classification techniques application, and evaluating performance measures for the classification of news articles.
We want to categorize the Bangla news from online Bangla newspapers. Also, we want to categorize the news headlines. There are various types of categories in a newspaper such as National, Politics, International, Opinion, Sports, IT etc. We want to find those categories from a newspaper.

2 Motivation

As we all know, there are several online news portals nowadays. Some online news portals show news on the basis of their categories. Some news Companies want to categorize the news based on published news in newspapers. So this work will be a solution for themselves. There are many works of categorization system of news headlines for another language. But there are a few works for the Bangla newspaper. So, our project is important to categorize the news of Bengali newspapers.
This project will help a lot to the people who are reading newspaper on regular basis.

3 Methodology

In our project, we used two model to predict the result of my project: These models are:

3.1 Model 1 Description : LSTM

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single data points , but also entire sequences of data.

3.2 Model 2 Description: GRU

GRU means Gated recurrent unit. GRUs are a gating mechanism in recurrent neural networks, introduced in 2014 by Kyunghyun Cho et al. The GRU is like a long short-term memory (LSTM) with a forget gate, but has fewer parameters than LSTM, as it lacks an output gate.GRU's performance on certain tasks of polyphonic music modeling, speech signal modeling and natural language processing was found to be similar to that of LSTM. GRUs have been shown to exhibit better performance on certain smaller and less frequent datasets.

4 Experiments

4.1 Dataset

The data was collected from various Bangla newspapers with scraping. There is more than one lac data in Author's dataset. They collected data from various newspapers like Bangladesh pratidin, Dainik juganttor, Daily Inqilab, Kalerkantho, and so on. These are the top visited newspapers in Bangladesh. They collected data from these newspapers and it helps this research

which categories data are mostly visited to the readers. They used Chrome Web Scrapper and python tools for scraping data from websites. There are three columns in Their dataset. These are Headlines, category and newspaper name. Their dataset is publicly available.

Total number of Headlines in Author's dataset is 103873.

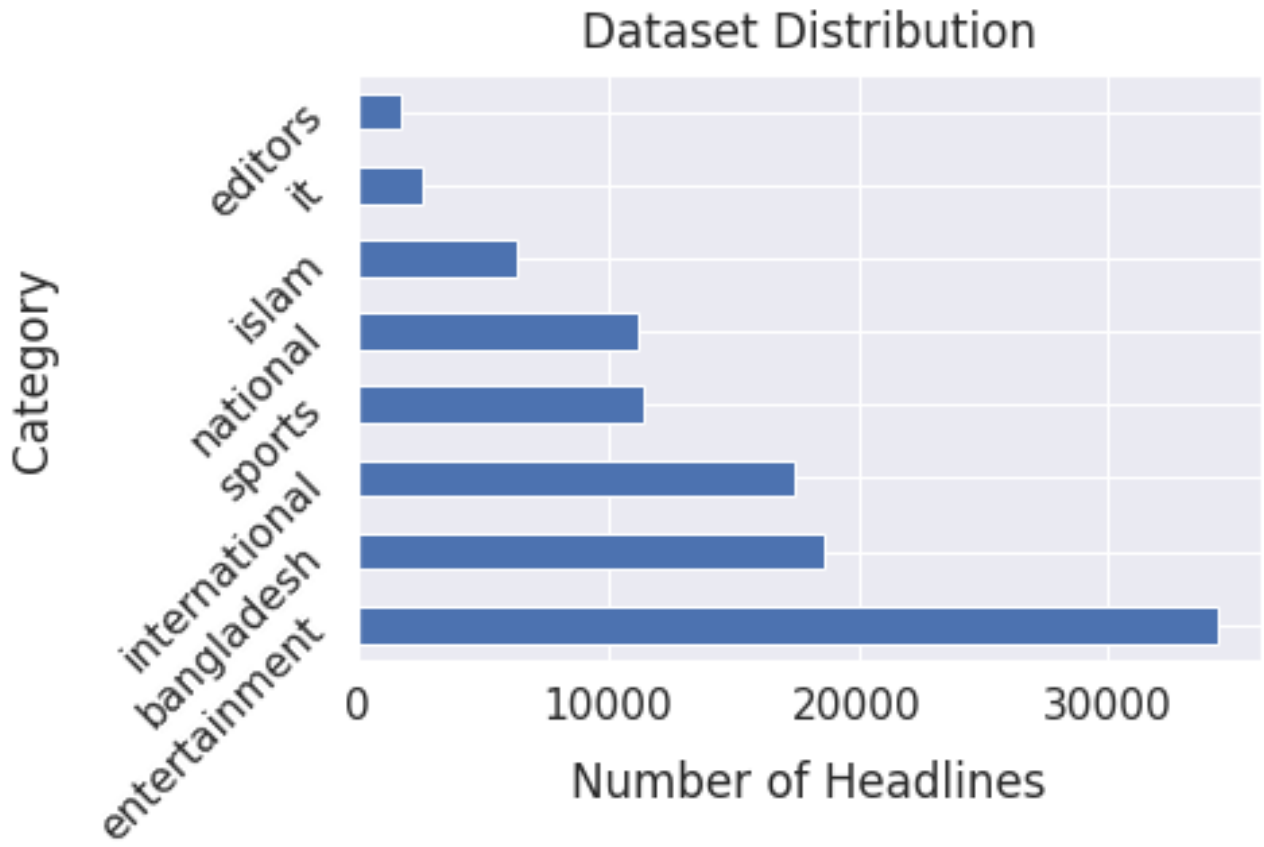

Figure 1: Headline Distribution Per Category

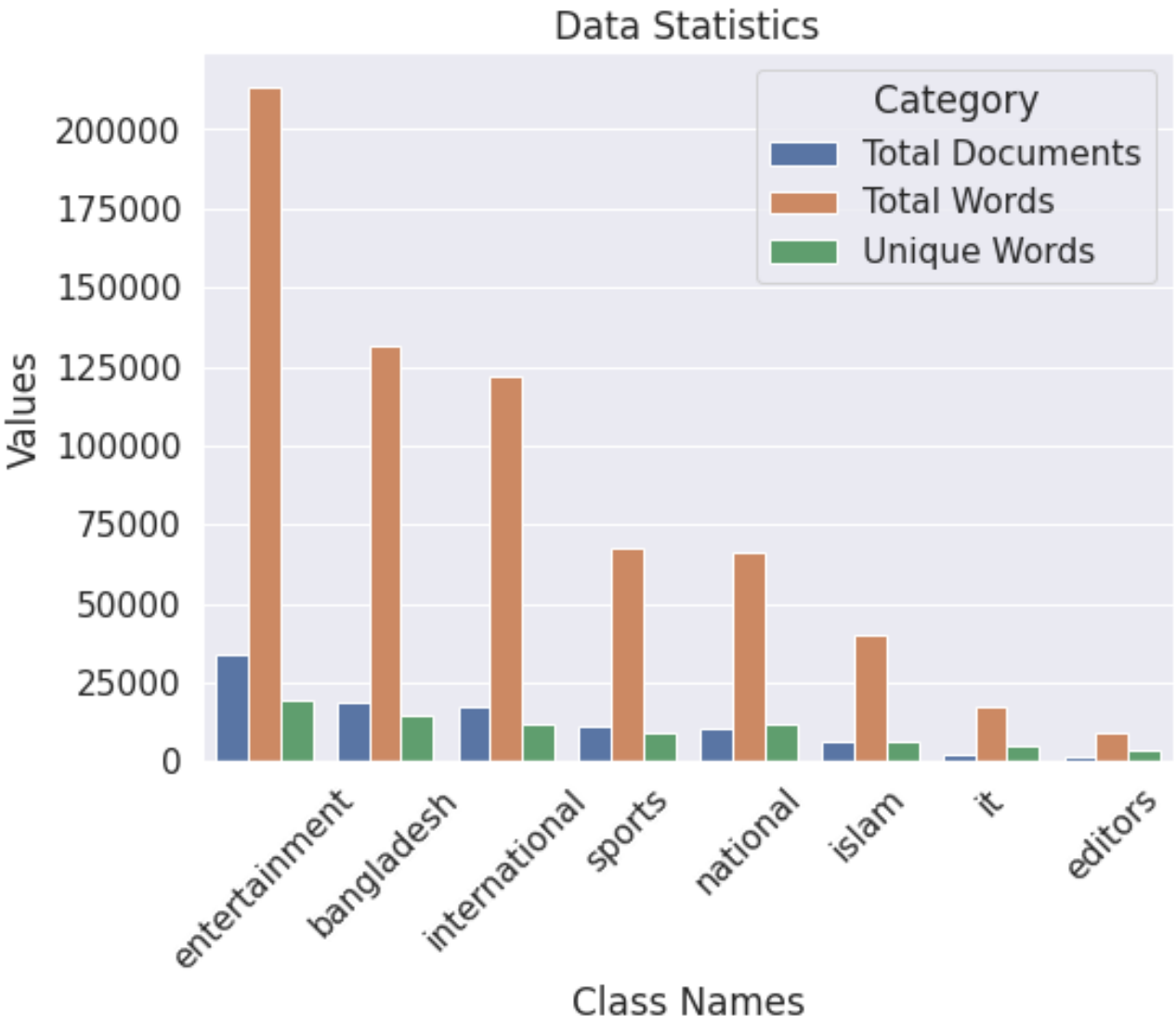

Figure 2: Word Distribution Per Category

Category	No Of Documents	No Of Words	Unique Words
Entertainment	33717	213598	19121
Bangladesh	18665	131368	14722
International	17440	121582	12032
sports	11438	67922	9048
National	10800	66501	12049
Islam	6382	40264	6508
IT	2560	17131	4842
Editors	1624	9160	3451

From the result table, We see Headlines and word distribution per category like number of headline, number of word and number of unique word etc.

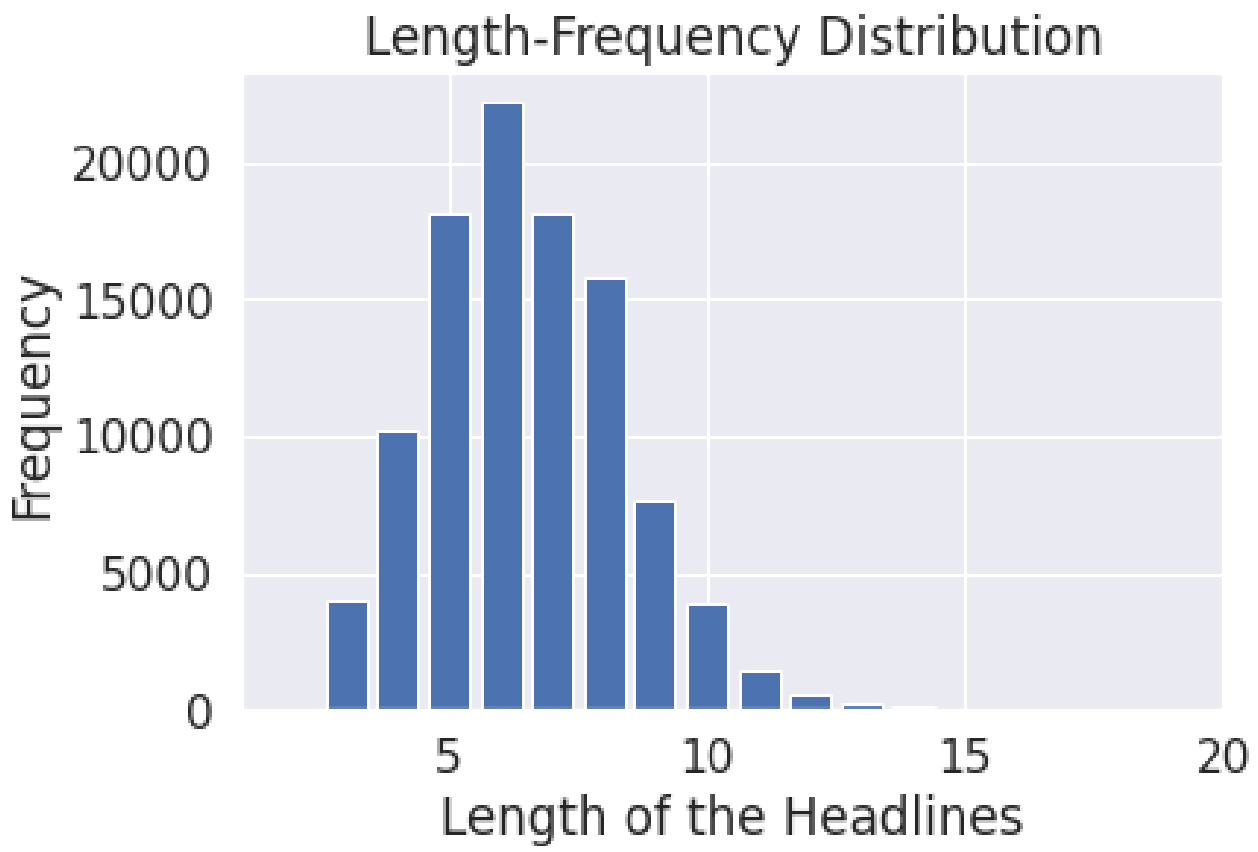

Figure 3: Length Distribution of Word

The Author's dataset have 103873 headlines and can categorized data into eight different category. These categories are:

1. Entertainment
2. Bangladesh
3. International
4. Sports
5. National
6. Islam
7. IT
8. Editors

There are 43661 unique words in their dataset. The Maximum length of headline is 19 words and minimum length of headlines is 3 words. There are on an average of 7 words per headlines.

Dataset Distribution:

Set Name	Size
Full	102626
Training	73890
Test	10263
Validation	18473

Figure 4: Dataset Distribution

4.2 Performance Metrics

4.2.1 Precision

Precision also known as positive predictive value. Precision value is defined by the ratio of correctly predicted positive observation to the total predicted positive observation. So, for our case precision value will be measure by the ratio between right prediction in a category and sum of Right Prediction and Wrong category Prediction .

Precision : (Right Prediction in a category) / (Right Prediction in that category + Wrong Prediction in that category)

4.2.2 Recall

Recall is the ratio of correctly predicted positive observations to the all observations(sum of predicted positive observation and false negative observation) in actual class. In this case Recall value will measure by the ratio between right prediction and total test data in that category.

Recall : (Correct Category prediction) / (Total test data in that category)

4.2.3 F1-Score

F1 Score is the balance between precision and recall. This score measured by the weighted combination of Precision and Recall. F1 Score is the ratio of twice product of Precision and Recall to the summation of Precision and Recall Score.

F1 Score : (2*(Precision * Recall)) / (Precision + Recall)

4.2.4 Accuracy

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions a model got right. Formally, accuracy has the following definition:

.

Accuracy : (Number of Correct Prediction) / (Total number of test data)

4.3 Results/Evaluation

4.3.1 Our Model Performance

We have used two models for predicting news headlines such as LSTM and GRU. We found different results from these two different models.

LSTM:
In this model, we have got 87.67% 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.

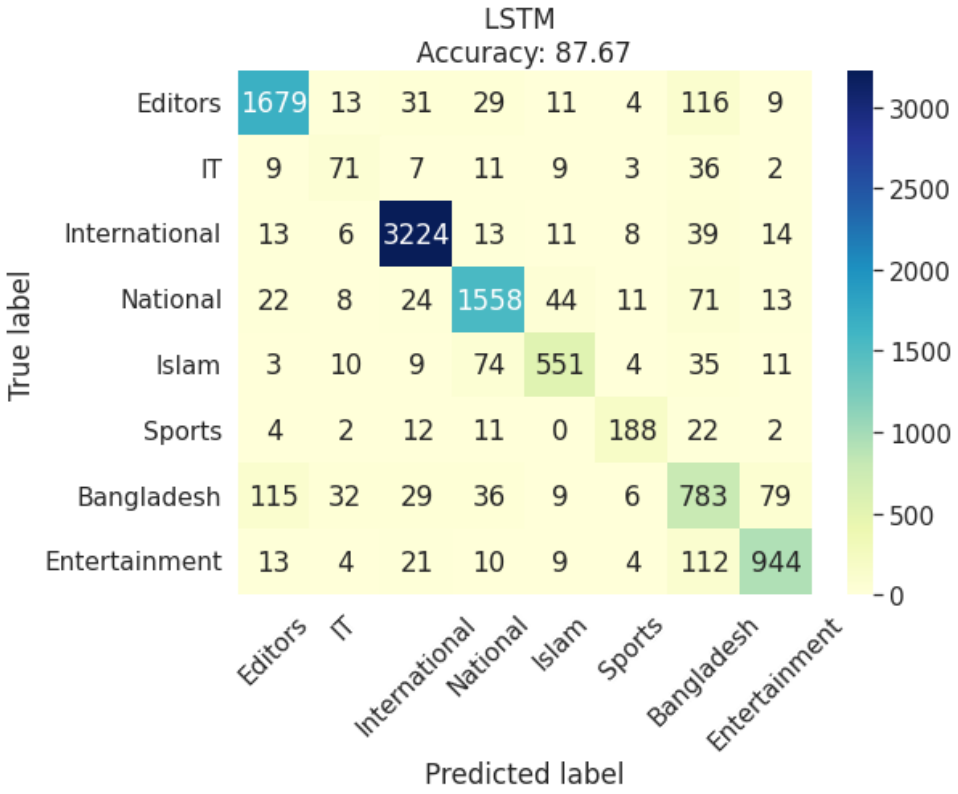

Figure 5: LSTM Model Accuracy

The accuracy, precision, recall, and f1-score result also demonstrate this issue.

	precision	recall	f1-score	support
Editors	90.37	88.74	89.55	1892.000000
IT	48.63	47.97	48.30	148.000000
International	96.04	96.88	96.45	3328.000000
National	89.44	88.98	89.21	1751.000000
Islam	85.56	79.05	82.18	697.000000
Sports	82.46	78.01	80.17	241.000000
Bangladesh	64.50	71.90	68.00	1089.000000
Entertainment	87.90	84.51	86.17	1117.000000
accuracy	87.67	87.67	87.67	0.876742
macro avg	80.61	79.51	80.00	10263.000000
weighted avg	87.92	87.67	87.76	10263.000000

Figure 6: LSTM Model (precision, recall, f1-score)

GRU:

In this model, we have got 87.96% 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.

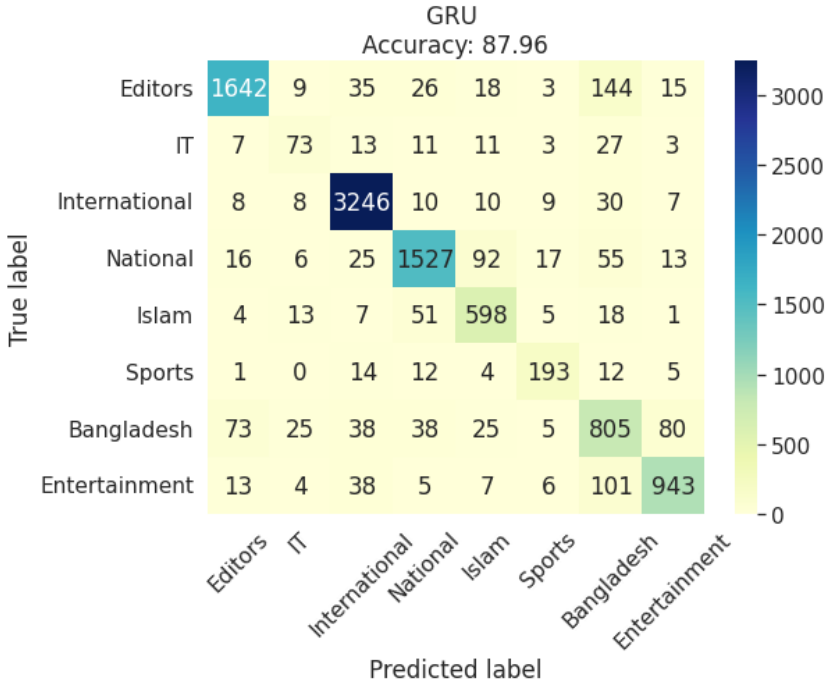

Figure 7: GRU Model Accuracy

The accuracy, precision, recall, and f1-score result also demonstrate this issue.

	precision	recall	f1-score	support
Editors	93.08	86.79	89.82	1892.000000
IT	52.90	49.32	51.05	148.000000
International	95.02	97.54	96.26	3328.000000
National	90.89	87.21	89.01	1751.000000
Islam	78.17	85.80	81.81	697.000000
Sports	80.08	80.08	80.08	241.000000
Bangladesh	67.53	73.92	70.58	1089.000000
Entertainment	88.38	84.42	86.36	1117.000000
accuracy	87.96	87.96	87.96	0.879567
macro avg	80.76	80.63	80.62	10263.000000
weighted avg	88.22	87.96	88.02	10263.000000

Figure 8: GRU Model (precision, recall, f1-score)

4.3.2 Result Comparison

Model	Our Accuracy	Actual Paper Accuracy
LSTM	87.67	82.74
GRU	87.96	87.07

Our model performs better than actual model.

4.3.3 Discussion about two results

We continuously changing our batch size since batch size is a hyper-parameter and we change it to 16,32,64 and 128 and analyze the differences in the performance matrix.For batch size 32 we get good result than the Author's in terms of precision, recall,F1-score and accuracy. Then we change another hyper-parameter Embedding Dimension(Embedding Dimension change the words to a fixed length Vector) 64 and 128 and analyze the performance. For embedding dimension = 128 we get slightly better result than the Author's. The final result of Author's in terms of accuracy is LSTM = 82.74 ; GRU

= 87.07 and Our result in terms of Accuracy is LSTM = 87.67 and GRU = 87.96.