UTILIZING MACHINE LEARNING AND DEEP LEARNING FOR CATEGORIZING BENGALI NEWS HEADLINES

BY

MD. SHAIKH AHEMED SHOVON
ID: 232-25-042

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science and Engineering

Supervised By

Dr. Md. Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Fizar Ahmed
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2025

APPROVAL

This Project/Thesis titled **"UTILIZING MACHINE LEARNING AND DEEP LEARNING FOR CATEGORIZING BENGALI NEWS HEADLINES"** submitted by **Md. Shaikh Ahemed Shovon**, ID No: **Student ID:232-25-042** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **11-01-2025**.

BOARD OF EXAMINERS

Chairman

Dr. S.M Aminul Haque
Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Fizar Ahmed
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md Alamgir Kabir
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Mr. Sadat Hossain
Data Scientist
Risk Management Division,
BRAC Bank Limited

DECLARATION

We here by declare that, this project has been done by us under the supervision of **Dr. Md. Zahid Hasan, Associate Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Md. Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Fizar Ahmed
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Md. Shaikh Ahemed Shovon
ID: 232-25-042
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Dr. Md. Zahid Hasan**,
Associate Professor, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning and Deep Learning" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. Sheak Rashed Haider Noori, Professor and Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

This study focuses on the application of natural language processing (NLP) in text classification, specifically for Bengali news headlines. Bengali, like many other languages, has seen increased attention in this field, with a primary focus on categorizing unlabeled news items into categories such as national, international, IT, and others. The growing popularity of Bengali news portals and the accessibility of online news make this a relevant area of research.

The proposed technique involves preprocessing steps, including tokenization, removal of numbers, special characters, and stop words, with a manually curated stop-word list to enhance performance. The study emphasizes the importance of stop-word elimination in feature selection. The methodology concentrates on classifying Bengali news headlines into eight distinct categories using machine learning models. Data is collected, preprocessed, and divided into training and testing sets.

The GRU (Gated Recurrent Unit) model demonstrated the best performance among the tested algorithms, achieving an accuracy of 84%. This result highlights the potential of machine learning techniques in effectively classifying Bengali news headlines. The study provides insights into the preprocessing methods and model selection processes that contribute to high classification accuracy, paving the way for further advancements in this area.

TABLE OF CONTENTS

LIST OF FIGURS

Chapter 1

Introduction

1.1. Introduction

Different methods assist the NLP system in understanding text and symbols. The act of categorizing a text using a specific set of terms is referred to as text categorization [1]. Text classification, also referred to as text categorization, involves sorting articles into predetermined groups. The task involves classifying free-form text into predefined categories. It offers theoretical perspectives on collections of documents and has practical uses [3]. It allows users to search within specific categories, making it easier to find information faster instead of searching through the entire information space. When the volume of information is excessive, the importance of organizing text becomes increasingly clear. Multiple researches have been conducted on news headline classification systems in different languages. However, there are only a few articles for the Bangla newspaper. Therefore, we developed a method for classifying news in Bengali newspapers. This study will assist in the creation of a self-sufficient system through the implementation of machine learning-based classification algorithms. In these methods, classifiers are generated or taught using a set of training documents. After that, the classifiers that have been trained are utilized to categorize documents into the correct categories. We chose to focus on online news as there is a lack of effective search features and visualization tools on existing news websites for evaluating data and trends amidst the vast amount of information on the internet. The continuous publication and citation of news data heightens the urgency of the issue. This led to the development of a system that serves two types of users: those interested in reading news stories by category, and those interested in examining data to identify trends in news.

1.2. What is headlines Categorization

In the field of data mining, there has been a recent focus on "Text Mining" as many researchers are conducting studies in this area. The process of extracting pertinent and precise data from extensive digital text is known as text mining. The significance of this subject in various industries illustrates its importance; for instance, it plays a crucial role in machine learning by utilizing methods of knowledge discovery to create logical guidelines for text classification. We will now create a model that categorizes news headlines. This will categorize the type of news headline.

1.3. Historical Background

1.3.1 Researchers who use real data gain advantages from classification methods. During a time with limited technological resources, scientists carried out some of the most adventurous research in history. Some researchers have found machine learning classifiers effective, while others have gained RNN access. This part examines previous research with high accuracy levels on the classifiers we have used, serving as an inspiration source.

1.3.2 Earlier Research

Yang Li created an SVM KNN method to classify short content [2]. CNN, SVM, NB, RNN, and LSTM machine learning classifiers were utilized. Ultimately, the SVMCNN classifier delivered better outcomes. By employing the SVM KNN technique, they reached results with an accuracy level of approximately 90%. Another important researcher, Tej Bahadur Shahi, made forecasts for automatic classification of articles in Nepal. He finished her investigation to choose a classifier model and artificial neural network. Machine learning classifiers such as SVM and Naive Bayes make use of multi-layer connectivity. The neural network, however, is currently in a slightly uncomfortable situation. During the process, 74.65% of Nepali news text categorization favored SVM, including RBF. Nevertheless, with an efficiency rate of 73 percent, the neural network is placed in second position in terms of ranking. Nepali news text categorization dataset consists of 4964 data points covering 20 various types of news. All types of deep learning models, like neural networks, require a large amount of data that has a high numeric value. Pranshengit Dhar and Md. are categorizing Bangla news headlines. Zainal Abedin utilized top machine learning principles [6]. SVM, Naive Bayes, and Adaboost were utilized as machine learning classifiers. They achieved an accuracy rate of around 81%. Sheikh Abujar suggested a Bengali news multi-classification system based on neural networks that achieved similar performance [7]. They create more than 86,000 news headlines. They utilized SVM, NB, Random Forest, Logistic Regression, and Neural Network as machine learning methods. Using Neural Network techniques, they achieved an accuracy of approximately 90%. Bjorn Gamback's primary focus was on text categorization to identify hate speech [8]. He desires to replicate the convolutional neural network. Using CNN assistance, they achieved an accuracy rate of 86.68 percent. They employ an alternative approach.

1.3.3 Recent Research

The primary goal of studying emotions is to categorize them as either positive or negative in order to distinguish between different parental attitudes or characteristics. The aim of this research is to enhance customer reach, earnings, and reputation. Methods, along with different sectors such as finance, economy, and spam detection in stock markets, buying and selling products, and various other industries. Effective intuition analyses can have a major influence in a variety of areas such as policy, governance, organization, campaigns, and enterprises, due to their ability to prompt swift reactions and allow individuals to benefit from necessary behavior or decision-making. Cost-effectively, neural networks can be obtained. A large number of emails, comments, and assessments adding up to thousands. Text categorization methods need to be expanded to cover businesses of all scales. Organizations need to be aware of and promptly address several crucial situations in order to act effectively. In order to quickly recognize important features, computer information retrieval must frequently imitate the designer's labeling in real-time. The concept of text categorization is well-established in the realm of natural language processing. Work on the Bangla text is ongoing. There are many different ways that online news is classified in this sector. In this age of online news sources, individuals rely on this matter. The proposed research, which focuses on the Bangla language, aims to achieve this classification. Some Bangla datasets utilize study materials found in our literature survey section. Our hybrid modeling approach achieves greater success than other machine learning approaches.

1.3.4 State of the art technology

Large CNNs, or those with many deeper and closely linked layers, often perform the best . Due to the high computational cost of these structures, big CNNs are sometimes unusably sluggish, especially for embedded IoT devices. Recent studies have focused on ways to keep prediction accuracy high while lowering the computing cost of deep learning networks for use in common applications. We studied the accuracy and computational requirements from pertinent literature, including current updates of networks as indicated in Fig 1, in order to comprehend the application of the state-of-the-art CNN architectures in agricultural systems.

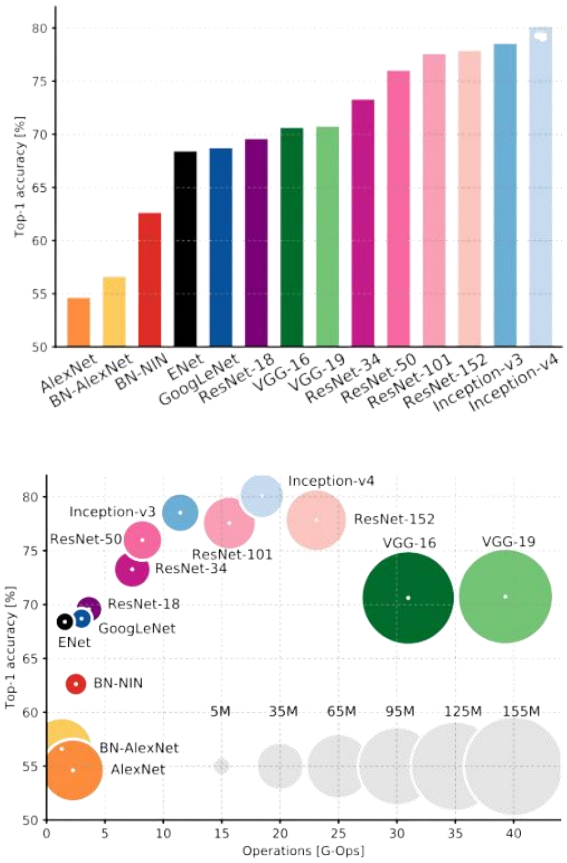


Fig. 1: Top-1 Accuracy vs the computational cost.

1.3.5 Future Scope of This Study

Although impressive progress in using cutting-edge CNN in agriculture in general, there are still unresolved issues with smart farms that future academics may examine. These fields can include interactive object identification, real-time image classification, and real-time image categorization. Modern CNN is a relatively new technology, which helps to explain why the study's findings about its usage in smart farms were not as strong. It's crucial to remember that models created using cutting-edge architectures have a proven track record of performing with greater precision.

1.3.6 Future Scopes

Neural Network reveals a field of study for Bengali researchers as a result of this multiclassification research. Others will be open to our suggested paradigm because text processing has recently gained popularity in the research community. Researchers

studying Bengali will also benefit from NN's automated word selection. In the event that the dataset grows in size, machines may likely offer more accuracy.

1.4 Limitation of the Study

With the use of a particular sort of dataset, classification is a system of kernels that enables an approach to anticipate its own choice. On our Bengali complicated dataset, we have employed neural network and machine learning classifiers including SVM, RF, LOGISTIC Regression, and NB throughout this study. As a result, by utilizing classifiers, the neural network produces a considerable output. Other classifiers provide results that are below average. To ensure a successful conclusion, we provided a sizable dataset. Before applying classifiers, the data needed to be preprocessed using an improving tokenizer and stopwords were removed from around 30000 of the datasets.

1.5 The Advantage over Traditional Method

The models used in earlier studies needed significant parameters and were substantially more complex. However, we are creating a much simpler model in this instance. Furthermore, our approach can quickly and accurately generate a good result even with uneven datasets.

1.6 Our Proposed Model

In this work we proposed a deep cnn model in our dataset and we gained a good result than other. We used two deep cnn. We used “RELU” and “sigmoid” as our model’s activation function. And we gained 84.01 in GRU and 83.42 in LSTM.

1.7 Objective of this Work

Our goal is to design a better model that can categorized headlines.

1.7.1 Primary objectives

First create a model that is get more accuracy than other previous model

1.7.2 Secondary Objectives

Second, create a model that is gain better perform than other previous ML model and others.

Chapter 2
DEEP CNN

2.1. Introduction

Deep learning is a computer modeling and machine learning technique that mimics how individuals learn. Data science, which also includes statistics and predictive modeling, includes deep learning as a critical component. Deep learning is helpful because it streamlines and accelerates the process for data scientists tasked with collecting, analyzing, and interpreting vast amounts of data. The most fundamental form of automation for predictive analytics may be thought of as deep learning. Deep learning algorithms are stacked in a hierarchy of increasing scale and abstraction as opposed to machine learning algorithms, which are linear.

2.2. CNN

An image may be inputted into a convolutional neural network (CNN), which can then prioritize and distinguish between various picture aspects. Compared to other classification algorithms, a Conv Net requires far less pre-processing. Furthermore, although basic procedures need filter engineering by hand, Conv Nets can learn these filters/characteristics with sufficient training. Conv Net's architecture was influenced by the way the visual cortex is set up, which is similar to how neurons are connected in the human brain. Individual neurons only react to stimuli in the Receptive Field, a small part of the visual field, which is made up of overlapping, comparable regions that make up the total visual field. Currently, CNNs are effectively used in three main methods for classifying medical images: 1) "CNN from scratch" training 2) utilizing "off-the-shelf CNN" features (without additional training the CNN) as supplemental information channels to already-existing hand-crafted image features, such as those used in chest X-rays and CT lung nodule identification; and 3) using CNN or other deep learning models to perform un - supervised from before the on organic or healthcare pictures as well as good on healthcare image features. To overcome the "curse-of-dimensionality" problem, a region - based segmentation

2.5D view re - sampling and an accumulation of random view categorization scores are utilized here. This allows for the acquisition of a sufficient quantity of training picture samples.

2.3. RELU

The activation function in a neural network is in charge of converting the node's summed weighted input into the activation of the node or output for that input.

If the input is positive, the rectified linear activation function, or ReLU for short, will output the input directly; if it is negative, it will output zero. Because a model that utilizes it is simpler to train and frequently performs better, it has evolved into the standard activation function for many different kinds of neural networks.

2.4. Sigmoid

A restricted, variational, real function called a sigmoid has a non-negative derivative at each point, is given for all real input values, and has exactly one inflection point. The same concept is described by the phrases "sigmoid function" and "sigmoid curve."

Chapter 3

Methodology

3.1. Introduction

Here is the method: Information is collected from various Bangladeshi newspapers. We used the Python module BeautifulSoup to scrape news from the website. After collecting data, we remove unnecessary symbols from datasets and then summarize them. This part includes data on the quantity of words, documents, and distinct words in each category. We determine the length frequency distribution using the pure datasets. Preparation of the datasets for the model is necessary. We utilized 80% of the data for training and 20% of the data for testing. Next, assign labels to the data using a coded sequence. I trained the model using 10 epochs and a batch size of 64. Therefore, our model's data is prepared. We utilized two deep learning systems to predict news headlines and evaluate results. I trained the model using 10 epochs and a batch size of 64.

Therefore, we have prepared the data for our model. LSTM and GRU are two deep learning algorithms utilized for predicting news headlines and comparing the results. We found that these models yield accuracy, precision, recall, and F1 score. After that, the results will be evaluated.

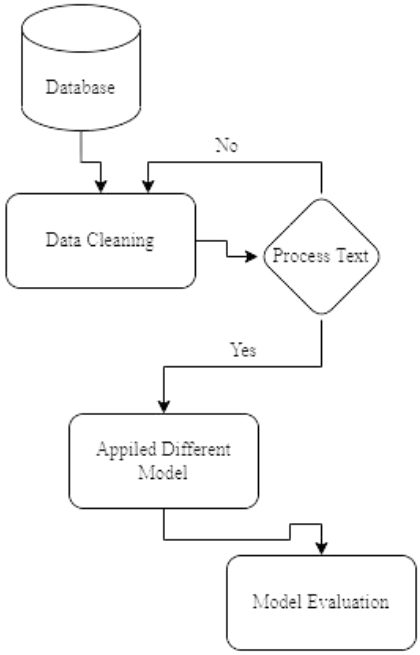


Fig-2: Model Architecture

3.2. Data Cleaning

Data can be gathered from a variety of sources. The headlines in a newspaper are divided into many categories. Data is gathered in real time from several Bangladeshi internet publications. Data is collected using scraping tools and technologies.

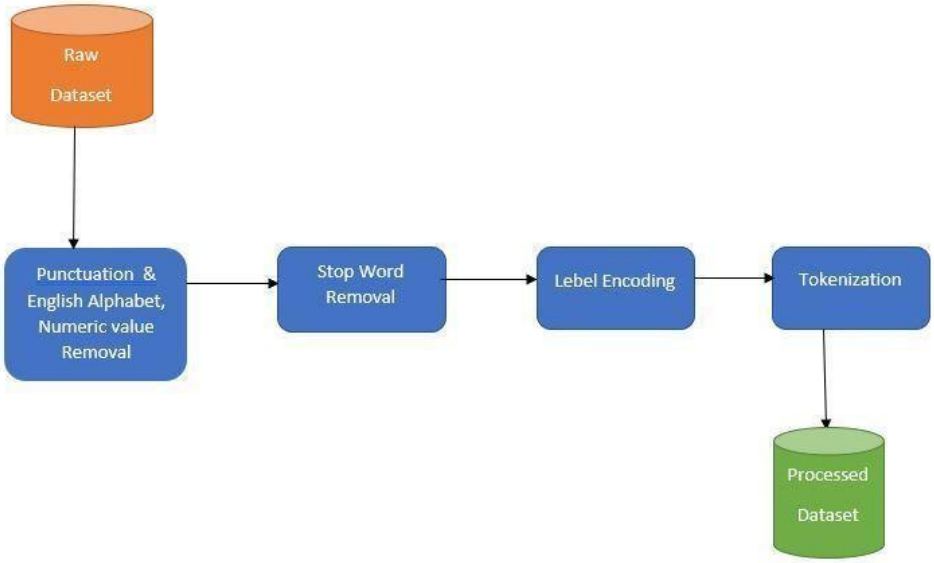


Fig-3: Data Preprocessing

3.3. Data collecting

Scraping was utilised to collect data from multiple Bangladeshi newspapers. We have more than one million records in our collection. We collected data from sources like Bangladesh pratidin [17], dainik juganttor [18], daily inqilab [19], kalerkantho, and similar publications. The newspapers are the most popular in Bangladesh. The data we collected from these newspapers helps identify the types of data that readers access most often. We used Chrome Web Scrapper and Python tools to gather data from websites. Our dataset consists of three columns. This includes the news titles, the section, and the publication's title. The information is accessible by the public.

The graph below shows the spread of headlines for each category. This data set is not evenly distributed.

The data is shown in the diagram provided.

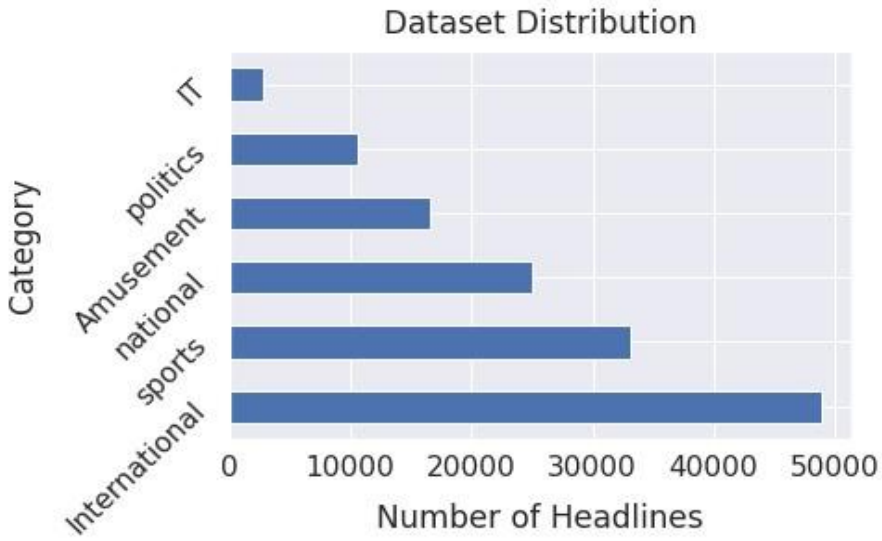


Fig-4: Dataset Distribution

3.4 Cleaning of data

Due to the brevity of the headlines, there is no need to remove the stopwords [20]. Regular expressions are employed for removing unnecessary data from our sample. The data sample will appear once it has been cleaned.

Original: ক্ষমা চেয়েও মুক্তি পেলেন না পরিচালক গাজী মাহবুব

Cleaned: ক্ষমা চেয়েও মুক্তি পেলেন না পরিচালক গাজী মাহবুব

Original: ব্র্যান্ডউইথের ব্যবহার ৮০০ জিবিপিএস ছাড়িয়ে

Cleaned: ব্র্যান্ডউইথের ব্যবহার ৮০০ জিবিপিএস ছাড়িয়ে

Original: জামিনে মুক্তি পেলেন ছাত্রদল সভাপতি

Cleaned: জামিনে মুক্তি পেলেন ছাত্রদল সভাপতি

Original: দ. কোরিয়ায় ১০০টি খালি কফিন পাঠিয়েছে যুক্তরাষ্ট্র

Cleaned: দ. কোরিয়ায় ১০০টি খালি কফিন পাঠিয়েছে যুক্তরাষ্ট্র

Original: ফ্লোরিডায় হামলাকারী 'মানসিকভাবে অসুস্থ' : ট্রাম্প

Cleaned: ফ্লোরিডায় হামলাকারী 'মানসিকভাবে অসুস্থ': ট্রাম্প

Original: সেরাটা দিতে পারলে সিরিজ জিতবে বাংলাদেশ:মাশরাফি

Cleaned: সেরাটা দিতে পারলে সিরিজ জিতবে বাংলাদেশ: মাশরাফি

After we finish cleaning the data, we can select the correct length for the headline to ensure that all headlines are the same length. Figure 4 displays the longest, shortest, and mean lengths of headlines.

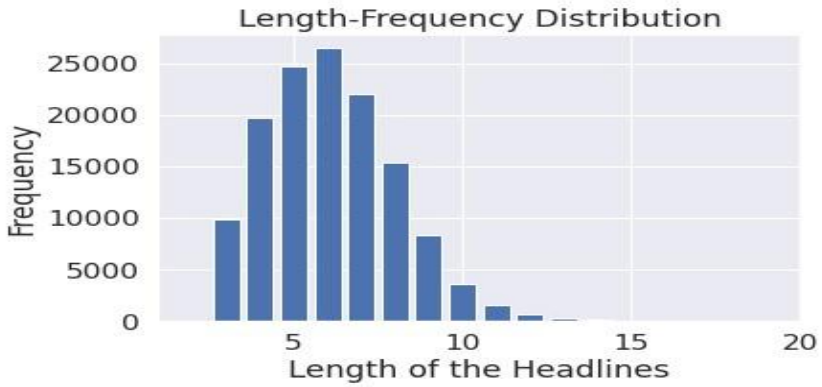


Fig-5: Length Frequency Distribution

Also, every category contains numerous terms. We select words from every category that are unique yet interconnected. The information in Figure 5 represents data statistics.

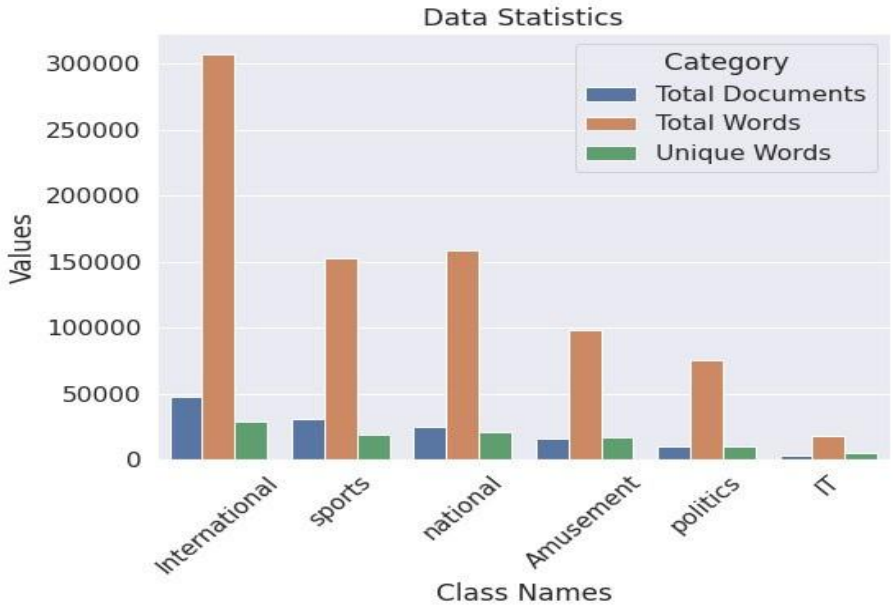


Fig-6: Dataset statistics

3.5. Data Preparation and Model Building

The encoded sequences represent the text data, with the sequences being the vector of an index number that contains words in each headline. Numeric values are also assigned to the categories. Following the preparation of the headlines.

====== Encoded Sequences ======

মোদির পাশে তৈমুর!
[4172, 2216, 6238, 301, 2629, 5925]

=====PaddedSequences======

মোদির পাশে তৈমুর!
[4172 2216 6238 301 2629 5925 0 0 0 0
0 0 0 0
0 0 0 0 0 0 0]

We add labels to each category after padding the sequence.

Chapter 4
Proposed DEEP CNN Model

4.1. GRU

Gated Recurrent Neural Network (RNN) applications that use sequential or temporal data have demonstrated success in a number of cases. They have been extensively used, for instance, in speech recognition, natural language processing, and machine translation. It has been successfully demonstrated that Lengthy Short-Term Memory (LSTM) RNN and the recently released Gated Recurrent Unit (GRU) RNN work well with long sequence applications. The gating network signals that regulate how the current input and past memory are used to update the current activation and create the current state are largely responsible for their success. In the learning phase, which includes the training and assessment procedure, these gates have unique sets of weights that are adaptively updated. Although these models facilitate effective learning in RNN, they also increase parameterization due to their gate networks. Consequently, compared to the straightforward RNN model, there is an additional processing cost. It should be noted that the GRU RNN only uses two gate networks, whereas the LSTM RNN uses three different gate networks. It is suggested that the exterior gates be cut down to a minimum of one with an initial assessment of long-term effectiveness.

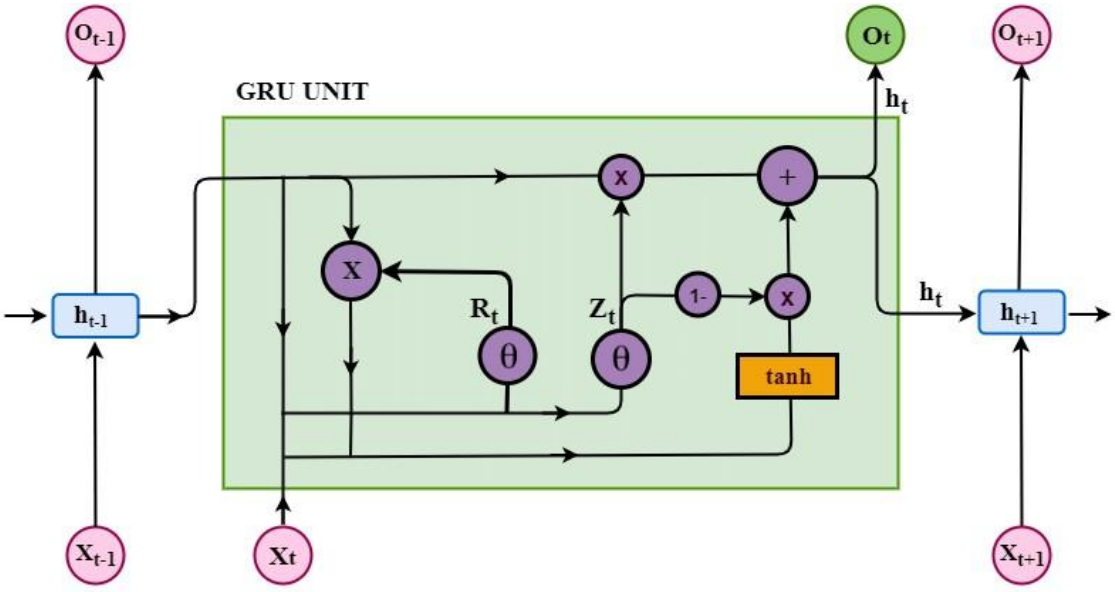


Fig-7: GRU Architecture

4.2. LSTM

RNNs have been created to manage time sequence data due to the recurrent architecture of the network. However, when there is a significant distance between the unit with the data and the unit where RNNs have difficulty learning to connect the information due to the gradient issues, especially with more information, problems can arise. Therefore, through the inclusion of three gates (input gate, forget gate, and output gate), Long-Short Term Memory (LSTM) Networks aim to improve upon the basic RNN structure. RNN and LSTM have shown their success in processing time sequence data in various fields such as action recognition, speech recognition, language translation, and image recognition.

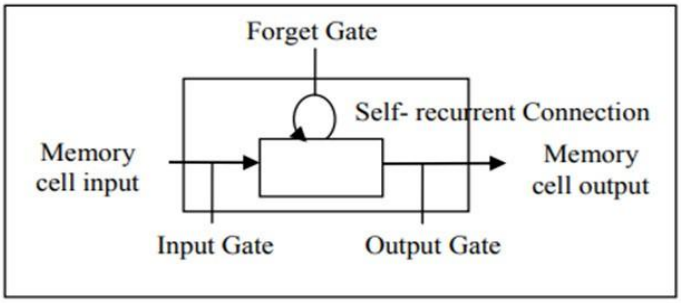


Fig-8: LSTM single cell structure

4.3. Accuracy & Precision

We have acquired a sense of measuring through millions of years of living because measurement is necessary for us to comprehend the outside world. Tools that supply scientists with a quantity are needed for measurements. The issue here is that there is some uncertainty in every measurement made with any measuring device. Error is the term used to describe this uncertainty. When collecting measures, accuracy and precision are two key things to keep in mind. These two phrases describe how closely a measurement resembles a standard or recognized value. Let's study more about precision and accuracy in this essay. In our machine learning approach the main algorithm setup was manufactured by the base of the random forest classifier. The Random Forest (RF) is a collection of separate decision trees. The "Gini index" of each branch is used to determine which choice line is more likely. Formula was used to obtain this index.

$$GINI = 1 - \sum_{l=1}^{c} Pi^2$$

In this case, c represents for the total class labels, while pi stands for the probability of a 10-th subclass. We selected 100 trees from the forest where its "gini" scale is employed to determine the degree of split. The nodes are divided if there is least 2 internal nodes and each internal node takes into consideration all system properties.

Chapter 5
Discussions and Conclusions

5.1 Result & Discussions

For forecasting news headlines, we employed Machine learning, LSTM and GRU. These two separate models had varied outcomes. The accuracy of models is discussed in Table 1. The GRU Model is more accurate. GRU produces superior results. Bidirectional model and soft-max activation function are both employed. The higher the score, the more data is tightly categorized. We used the GRU Model to properly classify the categories.

Table-1: Model accuracy

Model	Accuracy
LR	64.45%
MNB	61.33%
SVM	65.29%
RF	65.42%
GRU	84.01%
LSTM	82.74%
Bi-LSTM	83.42%

In the consistency of our model's evaluation, the highest accuracy and performance was acquired by the GRU which was 84.01%. In cases of ML approaches the Logistic Regression got 64%, the multinomial naïve bias achieved 61%, Support Vector Machine got around 65.29% and the Random forest was able to get the highest among the ML approaches consists of the 65.42%. And then if we focus on the deep learning approaches we can contemplate the variation of the GRU which stands at highest of 84% of total accuracy. As the LSTM has more subtle and linear approach it were able to acquire around 82% and the Bidirectional LSTM approach was withered a little better with 83% then the traditional LSTM approach.

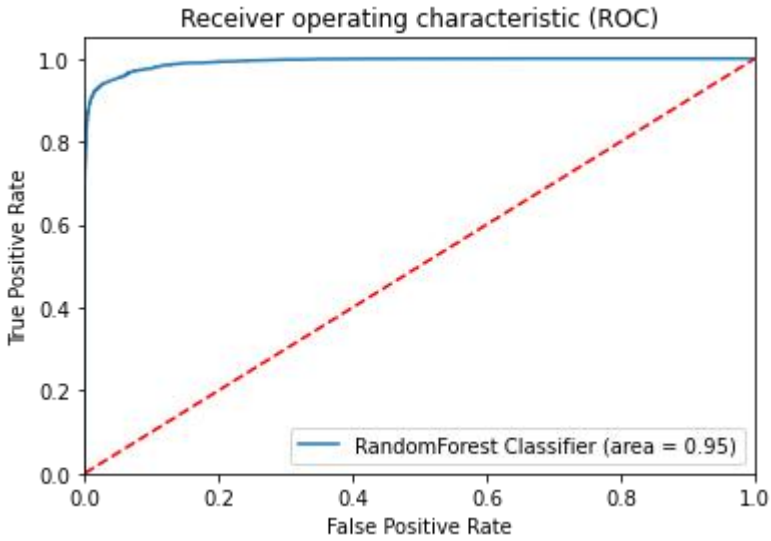


Fig-9: Random forest ROC curve

Fig-10: Confusion Matrix

	precision	recall	f1-score	support
Amusement	84.59	88.25	86.38	1617.000000
IT	64.04	45.45	53.17	286.000000
International	89.19	90.11	89.65	4852.000000
National	68.82	72.81	70.76	2398.000000
Politics	72.35	56.36	63.36	1054.000000
Sports	91.38	93.08	92.23	3065.000000
accuracy	83.80	83.80	83.80	0.838005
macro avg	78.40	74.34	75.92	13272.000000
weighted avg	83.58	83.80	83.56	13272.000000

	precision	recall	f1-score	support
Amusement	83.10	90.91	86.83	1617.000000
IT	59.77	55.59	57.61	286.000000
International	89.82	88.95	89.39	4852.000000
National	68.73	73.23	70.91	2398.000000
Politics	72.49	59.49	65.35	1054.000000
Sports	93.59	91.97	92.78	3065.000000
accuracy	83.99	83.99	83.99	0.839888
macro avg	77.92	76.69	77.14	13272.000000
weighted avg	84.04	83.99	83.92	13272.000000

Fig-11: Performance measurement

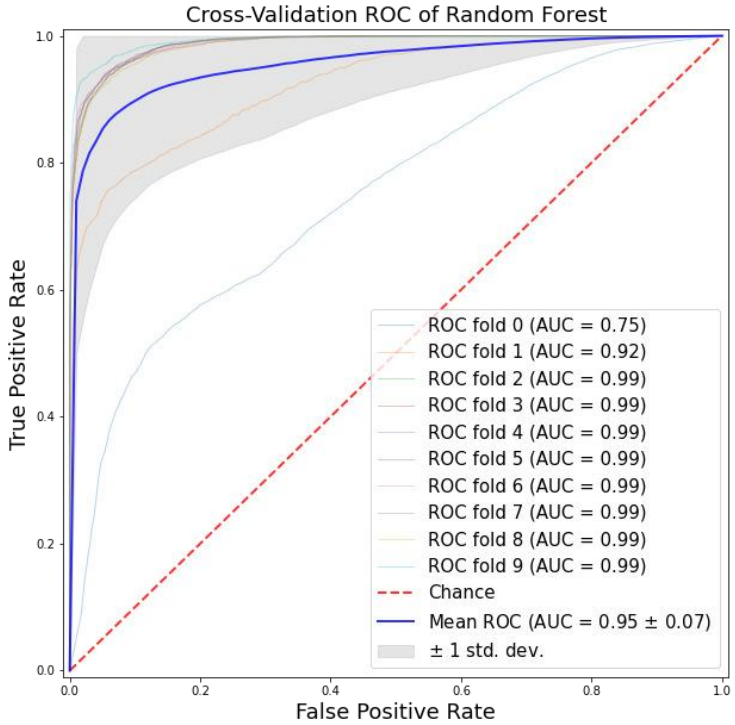
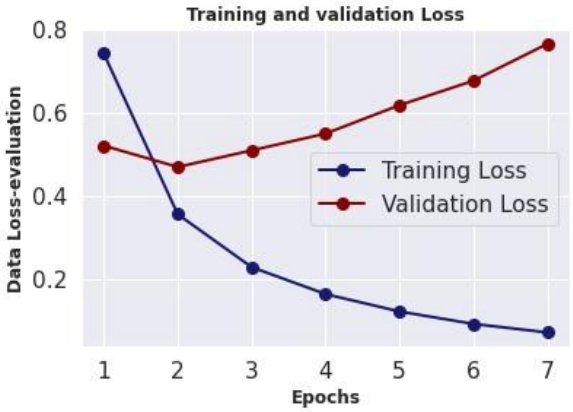
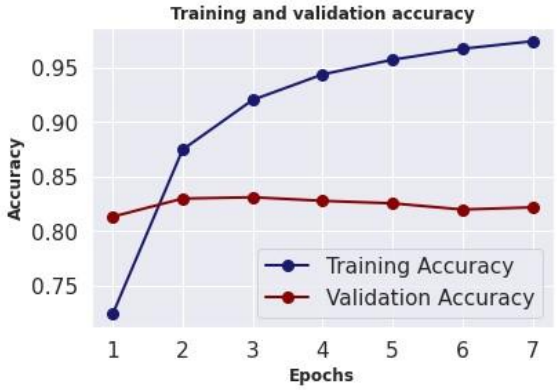
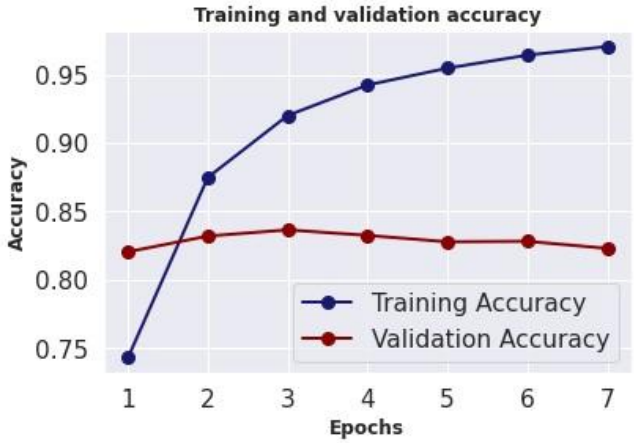
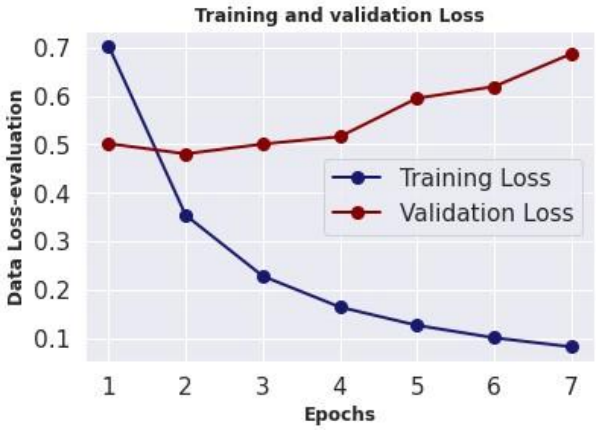


Fig-12:Validation ROC curve


Training and validation Loss
Data Loss-evaluation
Epochs
Training Loss
Validation Loss


Training and validation accuracy
Accuracy
Epochs
Training Accuracy
Validation Accuracy


Training and validation accuracy
Accuracy
Epochs
Training Accuracy
Validation Accuracy


Training and validation Loss
Data Loss-evaluation
Epochs
Training Loss
Validation Loss

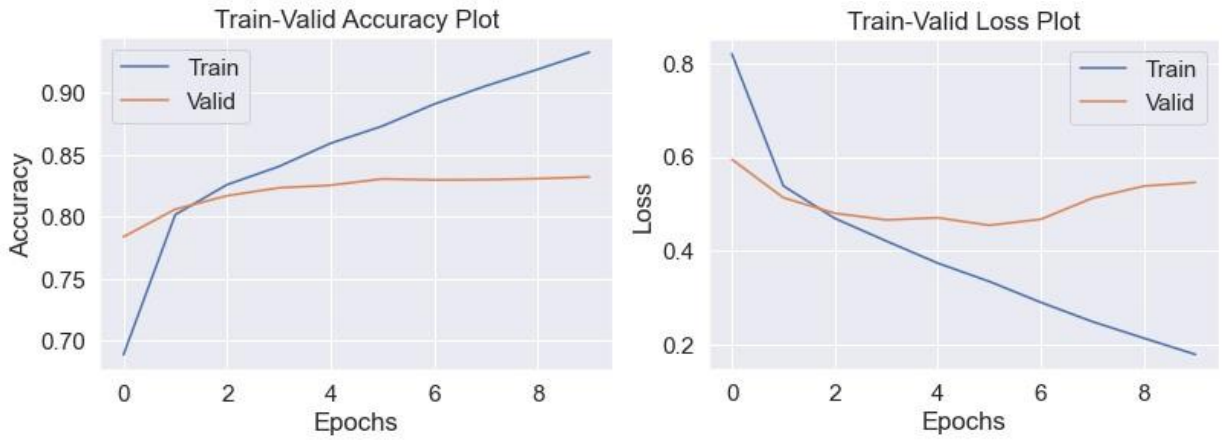


Fig-13: Bi-LSTM, GRU & LSTM train and validation accuracy and loss

5.2 Suggestion for future Work

In the examples provided, it is clear that our work encompasses a range of contrasting elements. The models and methods we utilized may form the base for our upcoming endeavors in natural language processing, machine learning, and deep learning studies. Developing Bengali language datasets has proven to be a significant challenge in this field, but we are committed to providing strong support to help advance research in computer science and bring about valuable societal progress.

5.3 Conclusions

This study created a model for news headlines using machine learning techniques. Classification of Bangla newspapers. Most studies in the literature reference another linguistic publication. GRU and LSTM are the top algorithms for creating an effective model in this classification method. The outcomes of the classifications are mostly consistent with earlier studies. Due to the utilization of two different methods for this classification, the outcomes could vary between each model. We have selected eight categories for news classification. The results do not depend on the categories. This method provides a more precise result with increased data, which includes both balanced and diverse data. Companies aim to categorize news based on the type of content that has been featured in the newspaper. Therefore, they could achieve the results they want. General consensus: more studies are required. This dataset is very small. Therefore, if we utilize multiple datasets, we can achieve better outcomes. Altering the attributes of the model resulted in different results. The result will vary as epochs change. Furthermore, the absence of the activation function in the models results in an impact. Multiple machine learning models are accessible. Different outcomes are generated by multiple models.

REFERENCES

[1] Meparlad, Understanding Text Classification in NLP with Movie Review Example, AnalyticsVidhya, (2020). [2] Shahin, M. M. H., Ahmmed, T., Piyal, S. H., & Shopon, M. (2020, June). Classification of bangla news articles using bidirectional long short term memory. In *2020 IEEE Region 10 Symposium (TENSYMP)* (pp. 1547-1551). IEEE.

[3] Yang, Y., & Joachims, T. (2008). Text categorization. *Scholarpedia*, *3*(5), 4242.

[4] Hu, Y., Li, Y., Yang, T., & Pan, Q. (2018, November). Short text classification with a convolutional neural networks based method. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)* (pp. 1432-1435). IEEE.

[5] Stein, R. A., Jaques, P. A., & Valiati, J. F. (2019). An analysis of hierarchical text classification using word embeddings. *Information Sciences*, *471*, 216-232.

[6] Omidvar, A., Jiang, H., & An, A. (2018, September). Using neural network for identifying clickbaits in online news media. In *Annual International Symposium on Information Management and Big Data* (pp. 220-232). Springer, Cham.

[7] Cai, J., Li, J., Li, W., & Wang, J. (2018, December). Deeplearning model used in text classification. In *2018 15th international computer conference on wavelet active media technology and information processing (ICCWAMTIP)* (pp. 123-126). IEEE.

[8] Shahi, T. B., & Pant, A. K. (2018, February). Nepali news classification using naïve bayes, support vector machines and neural networks. In *2018 International Conference on Communication Information and Computing Technology (ICCICT)* (pp. 1-5). IEEE.

[9] Dhar, P., & Abedin, M. (2021). Bengali News Headline Categorization Using Optimized Machine Learning Pipeline. *International Journal of Information Engineering & Electronic Business*, *13*(1).

[10] Khushbu, S. A., Masum, A. K. M., Abujar, S., & Hossain, S. A. (2020, July). Neural network based bengali news headline multi classification system: Selection of features describes comparative performance. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE. [11] Al-Tahrawi, M. M. (2015). Arabic text categorization using logistic regression. *International Journal of Intelligent Systems and Applications*, *7*(6), 71.

[12] Zia, T., Abbas, Q., & Akhtar, M. P. (2015). Evaluation of Feature Selection Approaches for Urdu Text Categorization. *International Journal of Intelligent Systems & Applications*, *7*(6).

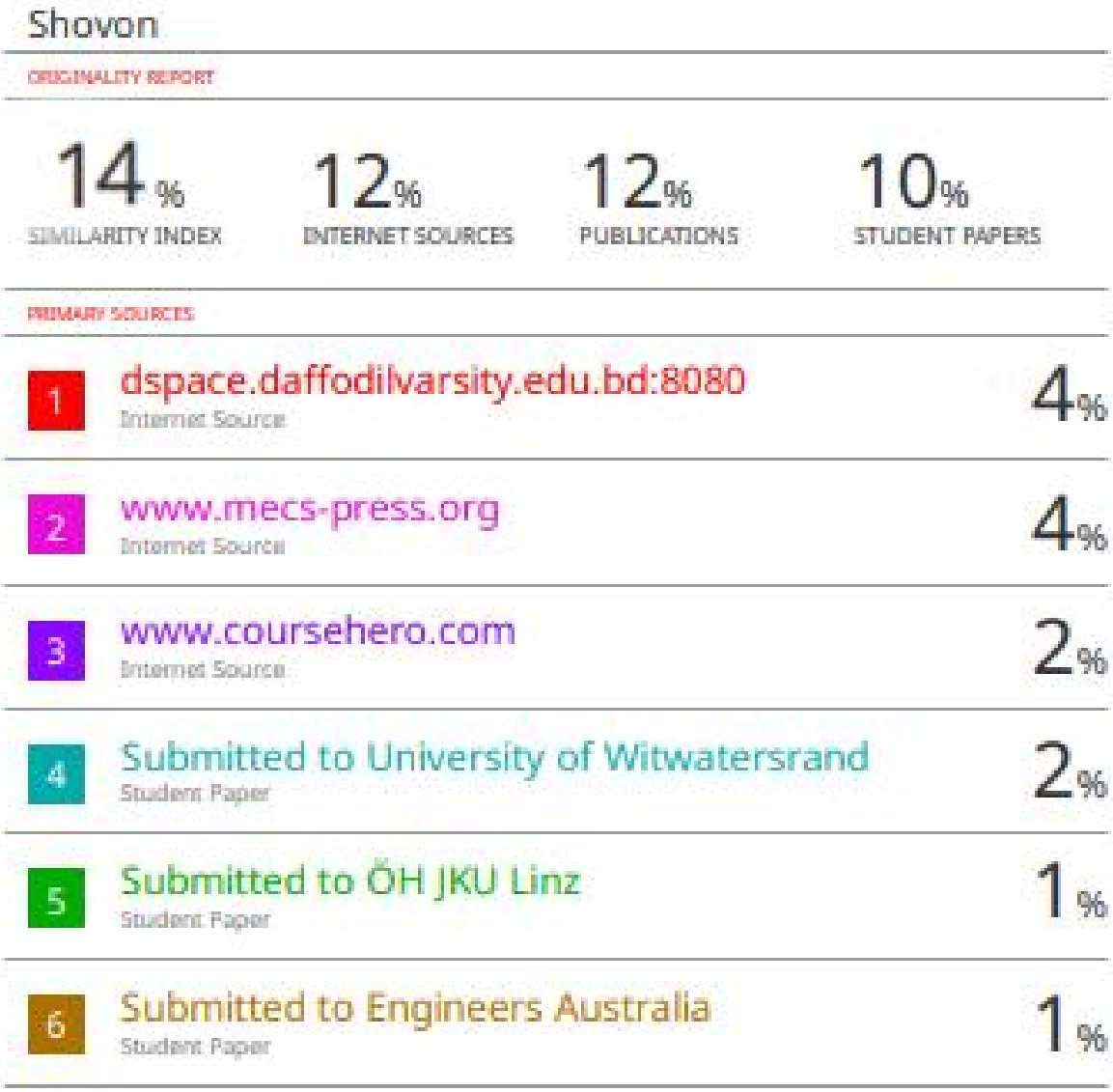
[13] Gambäck, B., & Sikdar, U. K. (2017, August). Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online* (pp. 85-90).

[14] Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4580-4584). IEEE.

[15] Kostadinov, S. (2017). Understanding GRU networks. Towards Data Science. *Towards Data Science, Towards Data Science*, *16*.

[16] Bangladesh protidin, https://www.bd-protidin.com (2021).

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



Shovon

ORIGINALITY REPORT

14%	12%	12%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	4%
2	www.mecs-press.org Internet Source	4%
3	www.coursehero.com Internet Source	2%
4	Submitted to University of Witwatersrand Student Paper	2%
5	Submitted to ÖH JKU Linz Student Paper	1%
6	Submitted to Engineers Australia Student Paper	1%