Multiclass Classification of Bengali Newspaper Article Using Transformer & Deep Learning Approaches

BY

MD. AHSAN HABIB
ID: 232-25-034

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

Supervised By

Dr. Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Abdus Sattar
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2025

APPROVAL

This Thesis titled **"Multiclass Classification of Bengali Newspaper Article Using Transformer & Deep Learning Approaches"**, submitted by **Md. Ahsan Habib**, ID No: **232-25-034** 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. Sheak Rashed Haider Noori, PhD
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md. Zahid Hasan, PhD
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Arif Mahmud, PhD
Associate Professor & Director MIS
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammed Nasir Uddin, PhD
Professor
Department of Computer Science and Engineering
Jagannath University

DECLARATION

I hereby declare that this research has been done by me under the supervision of **Dr. Naznin Sultana, Associate Professor, Department of CSE,** Daffodil International University. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

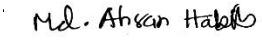
Supervised by:

Dr. Naznin Sultana
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Adbus Sattar
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

Submitted by:



Md. Ahsan Habib
ID: 232-25-034
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I express my heartiest thanks and gratefulness to Almighty Allah for Her divine blessing which makes it possible to complete the final year project/internship successfully.

I am really grateful and wish my profound indebtedness to **Dr. Naznin Sultana, Associate Professor**, Department of CSE, Daffodil International University, Dhaka, deep knowledge & keen interest of my supervisor in the field of Deep Learning to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

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

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

In this paper we introduce a large scale, structured dataset of Bangla news articles with 320k instances under several predefined classes (Science & Technology, International, National, Sports, Entertainment, Economy, Politics and Education) that aims to advance Bengali Natural Language Processing (NLP). Objective — to solve text classification problem for Bangla contents. A range of deep-learning models has been used for classifying the articles, where Bangla-BERT—a transformer-based model had attained an accuracy: 92% which was better than others. Other architectures (GRU, LSTM, CNN and a Hybrid Model) were also implemented and tested but Bangla-BERT outperformed with the highest accuracy. The present holistic dataset and the resulting insights on model performance allow a significant addition to available resources with Bangla NLP and an accurate benchmark for future works in this area. The implications of this work reach academics and industry; the Bangladeshi National Newspaper Organizations can use these models for efficient article categorization, and the natural language processing researchers are using an available dataset with insights on model effectiveness for Bangla text classification. This work represents a small step towards bridging the gap for NLP resources of Bengali language, and may pave the way for quantitative progress in automated language processing for Bangla.

.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

This research article is a groundbreaking work to classify Bengali news article with special emphasis on our daily use newspaper available in Bangladesh. Bangla has been ranked at 7th position among 100 most spoken languages across the world. In an article published on online portal, visual-content provider Visual Capitalist ranked Bangla in the top ten of the lists with 265 million native and non-native speakers across the globe.[1] It is yet to be a resourceful language in terms of various Natural Language Processing (NLP) applications. This paper address classifying Bengali news headlines which is very challenging since there are not many large annotated dataset in this domain.

Bengali has a rich history — Linguistic legacy found its significance in making the sacrifice of lives, for upholding the dignity of language which was evident in the historic 1952 movement[2]. And it is interesting to note that Bengali has attracted interest worldwide, with an increasing number of people learning the language as a foreign language. Bengali is a language that is spoken widely across the world but there are not sufficient text classification datasets for Bengali compared to some of other major languages, which hinders the development of NLP tools.

Text classification is an important process that helps organize and manage the information which is an integral part of applications like search engine, content management systems, news portals etc[3]. But Bengali is still lagging behind other languages regarding different fields of NLP, although for those languages with a larger dataset the development is enormous. In this paper, we fill in this gap by creating a dataset of Bengali news articles.

We crawled around 324k articles from different sources of well-known Bangladeshi newspapers and grouped them into eight categories: ScienceTechnology, International, National, Sports, Entertainment, Economy, Politics and Education to create the data set[4]. After data preparation, multiple transformer and deep learning models were trained in order to classify the documents into their respective classes. The models are GRU, LSTM, CNN, Hybrid (CNN+RNN), and Bangla-BERT. These strategies enabled that

study to tackle the challenges involved in classifying a low-resource language like Bengali, which requires larger datasets and suggests the requirement of sophisticated deep-learning model for more elaborate understandings of linguistic features.

Research methodology could be divided into four steps, starting with data extraction from trusted Bangladeshi newspaper like Prothom Alo, Jugantor, Ittefaq and POTRIKA available in kaggle dataset[4]. The data was preprocessed for tokenization, removing stop-words and then there were some filters to avoid punctuation marks / additional English words which lead to good machine-readable data. As observed, tokenization is pretty important since models don't work with text but numbers. The dataset was shuffled to ensure no data-aliasing effects would bias the data and then vectorized for model training as part of its pre-processing. A series of deep learning models such as GRU, LSTM and Bangla-BERT were trained with Bangla-BERT achieving the maximum accuracy proven to be effective for low-resource language processing. The classification accuracies with each model on the eight classes of assessment were then assessed.

This work provides an important dataset along with the trained models and the benchmark on it to further enrich Bengali NLP resources worthy of enabling future studies on text classification for Bengali language. And news organizations from Bangladesh will find useful AI solutions to help them categorize and optimize content available in the site. The outcome of this research can play an important role on Bengali text classification and integration to automated categorization as part of online news portal might help users in browsing the categorized articles. In particular, the conclusive observations from these studies include mentioning the promise of extending this work with larger datasets, data fairness and complex architectures that may provide better performance. In summary, this research helped in developing some technologies for Bengali NLP, specifically for content classification along with an opportunity to do more research on the Bengali language.

1.2 Motivation

Since my first semester of MSc in university, I have become passionate about applications developed from Machine or Deep Learning and this interest has never vanished. At first, the Deep Learning systems caught my attention but I quickly realized that this area has been established quite broad and so the motivation to build something into an existing well-studied problem wasn't strong compared to making a contribution towards work related to Bengali (my mother tongue). I came to this conclusion after reading many research articles in this domain of news classification for different languages and found none other than Bengali language news classification has been addressed. I feel this research would contribute work to the academics for developing Bengali language for amounting development in user experience and recognition.

I named the project —"Multi-class Classification of Bengali Newspaper Articles Using Transformer & Deep Learning Approaches". One thing to notice from the current trend is that technologies are getting advanced and how user experience is improved here and there in the whole world. Customers are growing accustomed to systems that instinctively provide relevant choices mapped out with their interest. So, to build a news recommendation system that people expect it must be the subject of independent decision making. That is what made me excited to do work that was research-based. This is the aim of such a goal which we will be focusing on in our study using deep learning and machine learning methods for it.

1.3 Rationale of the Study

Though huge papers have been published in the context of Natural Language Processing (NLP) and Text Classification, yet only a few studies were carried on Bengali news categorization[5]. Our work fills this gap by proposing a new method that combines data preprocessing techniques with the deep learning-based transformer model approach. The objective of this work is to make some powerful classification models using various algorithms which can help in making a strong classifier and it will add value in task from the domain of NLP.

NLP is a sophisticated method that allows us to analyze textual and speech records data to seek out the particular which means of a given context. This field is the programming of computers to simulate human thought processes, which utilizes natural human languages. It can be a text or speech output and input for the NLP systems which is in between human language and machine language. As one of the NLP features, Text Classification processes large sequences of information, identifying salient terms related to other information sequences for brief categorization and understanding[6]. The model learns to filter out the misleading information which improves and prices the accuracy and reliability through massive dataset training.

1.4 Research Questions

It was challenging to complete this research. In order to respond accurately and pragmatically to the sentiments and results involved, the researchers offer questions that can guide a response:

- RQ1: Can the large corpus of Bengali textual data be collected and then preprocessed to make it available for modeling?
- RQ2: Can this work help improve text classification systems?
- RQ3: Why in this context, hasn't been used a transformer model?

1.5 Expected Output

There are some of the key objectives defined under this section and these will be our main expected outcomes. Abstract—This project research to classify the Bengali news based on transformer or a whole, efficient approaches with the model build from training dataset.

- Bengali news can be categorized in different types.
- Such a frame can be a lot helpful to the news portals available on the internet.
- This framework could improve the experience of online news readers as well if adopted by all news portals.

1.6 Project Management and Finance

This research work did not receive funding from any individuals or organizations.

1.7 Report Layout

This report consists of six chapters, where we have explained all our work in as organized way as possible. For better understanding, we are providing a very brief summary below.

Chapter 1

This chapter has included all our introductions, motivation of work, our objectives, research question. In advance of our research, we have already highlighted the theory and work-related information antecedents to our research.

Chapter 2

In this chapter we have briefly discussed about the terminologies and related works. In addition, the summary of our research and area bounded in problem were explained. We closed this chapter by discussing the obstacles that we have encountered throughout doing this research.

Chapter 3

In this chapter, we described the methods of our work. Also touched on the technologies & equipment that we needed to use. We demonstrated our data collection procedure and data preprocessing. Finally, proposed our deep learning models, statistical examination, a blister of model details and structure, along with the depiction of Taxonomy of our model was all discussed.

Chapter 4

We have discussed the results and shown the comparison among the models which gives best performance for our classification. We have shown real-life prediction accuracy using a simple UI for each model to classify any bangla news article.

Chapter 5

We discussed chapter fives by the impact on society, ethical aspects and sustainability plan for understanding significance of our research work.

Chapter 6

Lastly, we covered future work and briefly discussed it. In this article, we explain a brief summary of study, recommendation, conclusion and implication for further study.

CHAPTER 2
BACKGROUND

2.1 Introduction

The primary task is multiclass classification which aims to assign each article into one of the possible categories. Natural Language processing (NLP) problems arise for Bengali text due to its rich morphology and are addressed in this paper. Deep learning models, especially transformer architectures (such as BERT [Bidirectional Encoder Representations from Transformers], CNN and RNN) can be used because they capture complex long-range dependencies in text by using self-attention mechanisms[7]. It was addressed the challenge of less resourceful for Bengali language also by pre-training on Bengali specific data then fine-tuning the model. It also uses pre-trained language models, which we adapted to the task via transfer learning suitable for obtaining better results with limited data. It involves applying word embeddings, tokenization, and other text preprocessing techniques to prepare the data, and using metrics like accuracy, precision, recall as well as macro and micro F1-score to evaluate model performance. This work attempts to improve the efficiency and efficacy of Bengali text classification by using combination of recent advanced deep learning methods and language resources appropriate for Bengali.

2.2 Related works

We are pleased to introduce the Potrika dataset, a valuable resource in the field of Bengali Natural Language Processing (NLP), filling the absence of large-scale datasets for text classification. Potrika consists of 665000 articles collected from six main Bangladeshi new portals in the period between 2014 and 2020 having eight categories namely National, Sports, International, Entertainment, Economy Education Politics and Science & Technology[8]. To address class imbalance, a balanced subset consisting of 40,000 articles per category (total: 320,000 articles) was constructed. We provide state-of-the-art results on a number of benchmark datasets as well as scale, category representation and balance over previously available Bengali resources namely, BERT or the works by Hossan and

Shahin. Potrika is suitable for various NLP applications like classification, named entity recognition and text summarization, making it a robust benchmark to build Bengali NLP models. This will be an important milestone for the development of Bengali NLP as it allows for data-driven solutions and advances research in low-resource languages[8].

Shortly after, the important papers were published in Bengali news classification, but they ended with proper comparison between other major languages and Bengali. Specifically, despite a well-resourced landscape for English NLP, Bengali does not enjoy the same level of availability of datasets and models. Related works generally revolve around machine learning and shallow types of deep learning with data sets being small sized or lacking diversity in terms of categories. Though a few approaches on text classification in Bengali only have been made [9,10] for example from BRAC University, CUET they followed the techniques of N-grams and convolutional neural network(NN-CNN) which produced good performance. However, these works highlight the lack of high quality balanced datasets and better models to reach satisfactory classification accuracy for Bengali, particularly for domain-specific tasks like news categorization.

Sentiment analysis and topic categorization among Bangla text using Natural Language Processing(NLP) tools, the literature on Bengali news classification indicates that great need of NLP in Bangla Text Analysis[11]. Unfortunately, there are not many resources and datasets available that can push the Bengali NLP forward. It is a remarkable attempt by some researchers (Potrika dataset) to come up with a well-balanced dataset based on categories such as National, Sports, Economy, Politics etc. Research implemented approaches may leverage either machine learning or deep learning techniques such as Logistic Regression, Support Vector Machine (SVM) and Gated Recurrent Unit (GRU). To increase accuracy, it often incorporates word embedding such as Word2Vec, FastText and TF-IDF. These approaches boost classification but limited annotated datasets exit along with the complexity of the Bengali language[12].

The existing literature on Bengali news classification indicates that the scope for performing Natural Language Processing(NLP) functions globally has much progressed

but NLP of Bengali language is a big area to explore due to lack of resources. Though, English has many datasets and pretrained models available for text classification like tasks, but Bengali just explores the tip of an iceberg with respect to scope as well as size of dataset. Prior work has mainly used conventional machine learning approaches (e.g., SVM, Naive Bayes), and neural networks such as LSTM, GRU[13]. But all above studies show us that they always needs well balanced and large datasets, also the models must be strong enough to achieve better performance in Bengali text classification specially for domain specific tasks such as news category.

Literature on Bengali news classification shows the disparity between NLP resources for Bengali and raises similar concerns, as most of the existing works utilize machine learning and deep learning approaches to address this. Although others have used techniques such as Support Vector Machine (SVM), Naive Bayes, LSTM, or GRU [13], the small and low diverse datasets do not allow better performance to be achieved with these models. Potrika dataset was built to fill this gap by enhancing the classification across multiple categories for a generalist newspaper to cater NLP tasks like topic classification and sentiment analysis. This work highlights the need for huge datasets and strong models to improve Bengali NLP capabilities.

This work establishes the state of the art in Bengali news classification and highlights the need for better decades-old language resources pertaining to low resource languages like Bengali. As much as work has progressed in English and other major languages, Bengali lags behind on account of lack of available datasets and models. Previous works often use machine learning methods like Support Vector Machines, Naive Bayes and Random Forests, as well as neural networks such as LSTM and GRU. Yet, often due to small, unbalanced datasets accurate. Recent resources, such as Potrika dataset, mitigate the aforementioned drawbacks by providing larger and comparatively balanced data sets making it possible to build better classification models especially in the context of Bengali news categorization.

2.3 The Problem's Scope

Although transformer models are known for its high accuracy and state-of-the-art performance (SOTA), especially in text classification tasks, they have not yet been extensively used in Bengali news classification. Current methods are mainly based on either conventional Machine Learning models or recurrent neural networks (like LSTM and GRU), which cannot effectively capture processed complex language features in very long sequences. We find a significant gap here which can be filled with transformer models such as Banglabert that already has shown the potential by outperformed in many different NLU tasks on Bengali including this paper for Bengali newspaper classification. The objective of the research is to familiarize transformer-based models for Bengali news classification, showing their capacity to deliver improved accuracy and efficiency while establishing a new guideline in relation to resources around Bengali NLP.

2.4 Challenges

This has been the biggest challenge of this project because first you need to, amass a lot of databases at scale and then clean it all up and structure it accordingly. This data cleaning has several steps, for stop words removal, English text and punctuations. For machine learning, the dataset needed some work in terms of tokenization, while for deep learning models we would need another layer here called vectorization. Batch size: As the dataset was quite large, it took much time in getting final model outputs for both machine learning and deep learning models. While a few datasets were available, the small size made it difficult to achieve accuracy thus, we needed to collect larger dataset for bengali text classification. This entailed creating everything from scratch, collecting and cleaning data from different publicly available datasets, designing and training all 5 models motivated solely by passion and a spirit towards pushing the frontiers of Bengali NLP.

CHAPTER 3
RESEARCH METHOLOGY

3.1 Introduction

In this chapter, I will describe the methodology I followed for this project. The operations that are carried out include downloading the data set, preparing the data, analyzing the data, choosing the model, training the model, validate the outputs and make the prediction. Both transformer and deep learning models were trained with caution. Data preparation is necessary because the dataset is one of the largest corpus in bangla newspaper.

3.2 Proposed Methodology

This study, on the other hand, has a methodology that includes data collection, data cleaning and pre-processing, tokenization and vectorization, model building and evaluation.

- **Data Collection:** The dataset we used here is from Potrika, which is the largest public Bengali news articles dataset available[6].
- **Data processing:** the data was grouped and analyzed right after it was collected from different sources. Selected dataset was manually reviewed and filtered of distorted, flawed or irrelevant data points, before we continued with the input.
- **Data Cleaning and Pre-processing**: At this stage, data was processed class-wise. English words, punctuation, special characters, whitespace and digits are removed to make the dataset trainable/testable. This pre-processing step was time-consuming because of the big dataset.
- **Building the model:** In this step, the data was tokenized and prepared to create a train-test split. We trained five models, including transformer and deep learning algorithms. This allowed us to compare the performance of the models in both training and testing stages.

- **Performance Evaluation:** In this part I share graphical and theoretical results, such as accuracy graphs, test loss, confusion matrices, etc related to the performance of each model
- **Conclusion and Future Work:** This section basically wraps up this study, but proposes plans for more progress in the future.

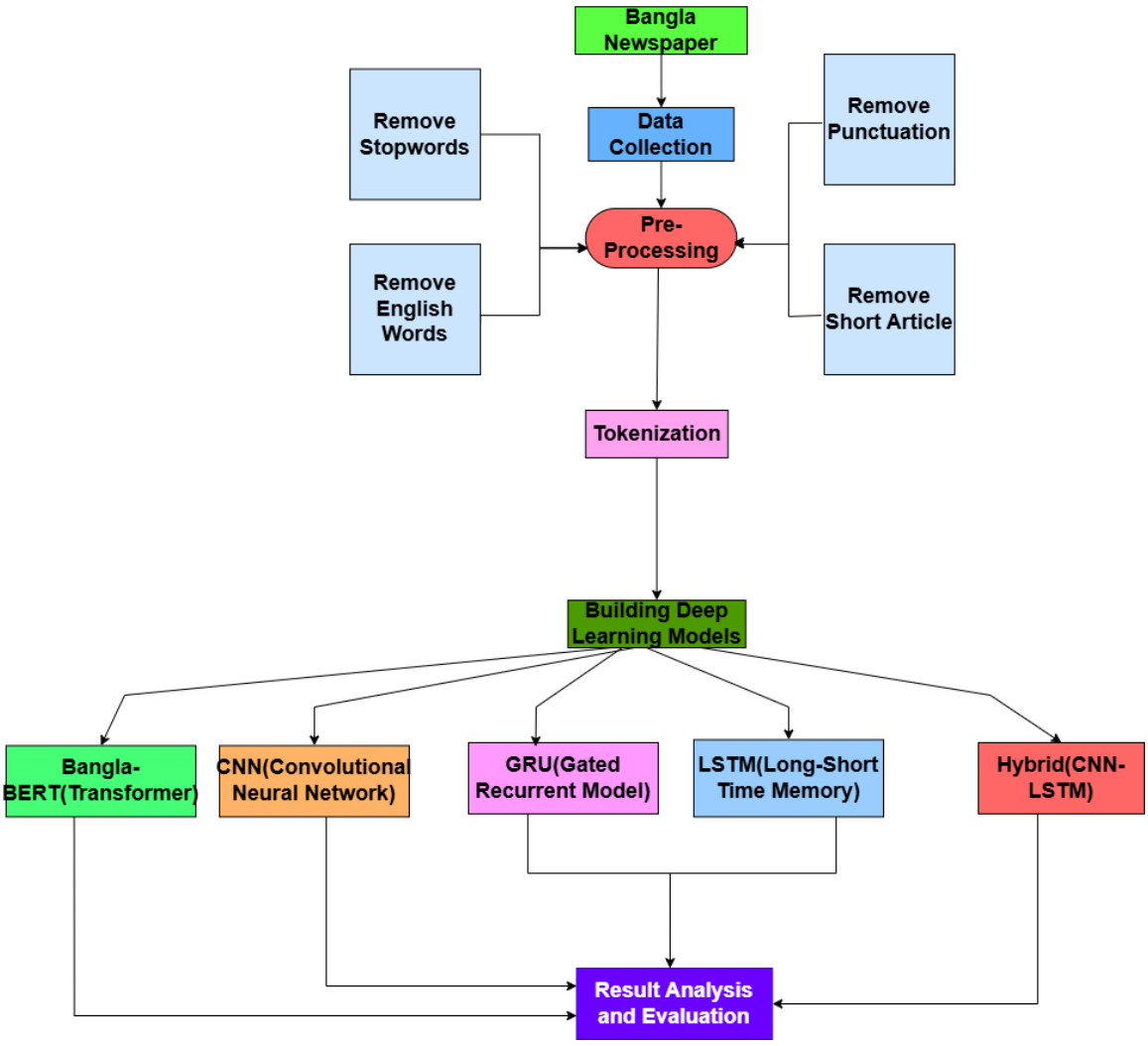


Fig 3.1: Workflow of Newspaper Article Classification

3.3 Data Collection Procedure

We have used Potrika dataset which is one of the largest Bangla newspaper datasets containing news articles collected from six major Bangladeshi news portals, namely Jugantor, Jaijaidin, Ittefaq, Kaler Kontho, Inqilab and Somoyer Alo. Our dataset from 2014 to 2020 consists of 664,880 articles divided into eight different categories National, Sports, International, Entertainment, Economy, Education and Politics and Science & Technology[6]. Potrika dataset was processed to remove distortions and irrelevant content, specifically preparing the data for NLP tasks like text classification. The author of the dataset applied data augmentation techniques to balance the dataset and create 40k articles for each category, so that machine learning and deep learning models could be trained on a balanced dataset for Bengali news classification. This dataset serves as a useful benchmark for Bengali NLP research and applications.

	article	category
0	সাভারের কবিরপুর বাণিজ্যিক এলাকার নিজস্ব ফ্যাক্...	Economy
1	\nরিজার্ভ চুরিচলতি বছরের সবচেয়ে চাঞ্চল্যকর ঘটন...	Economy
2	অর্থনৈতিক রিপোর্টার : এসএমই ফাউন্ডেশনের কনফারে...	Economy
3	ধীরে কমছে চট্টগ্রাম বন্দরের জট। টানা বর্ষণ, ব...	Economy
4	\n\nমোবাইলফোনে কথা বলায় এবার খরচ বাড়ছে। বাজেটে...	Economy
...	...	...
40174	\n\n পরবর্তী কর্মসূচি না দেয়া পর্যন্ত সারা...	politics
40175	\n\n প্রধানমন্ত্রী শেখ হাসিনার নামে ঢাকা বিশ্ব...	politics
40176	\n\n ভাই জিএম কাদেরকে দলের কো-চেয়ারম্যান করা...	politics
40177	\n\n তথ্যমন্ত্রী ও জাসদ সভাপতি হাসানুল হক ইন...	politics
40178	\n\n বিএনপি নেতৃত্বাধীন ২০ দলীয় জোটের বৈঠক হবে...	politics

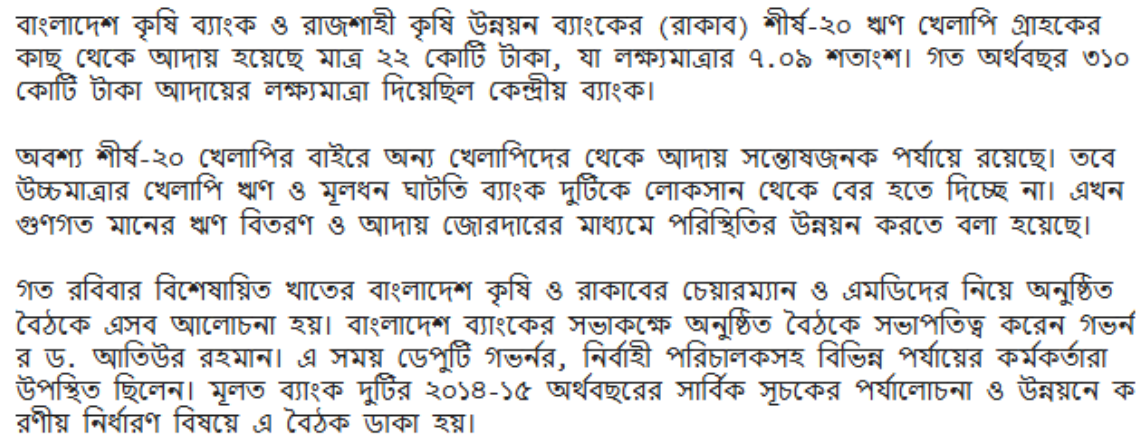
329110 rows × 2 columns

Fig. 3.2: Total number of articles in our Potrika dataset

3.4 Data Pre-processing

In this research, a lot of data preprocessing has been done to optimize the Potrika dataset for efficient machine learning and deep learning model training. The preprocessing steps included stop word removal, data cleaning, tokenization, and vectorization.

- **Cleaning the Dataset:** We filtered out any unnecessary parts of the dataset such as english words, punctuation marks, symbols or special characters, spaces and numeric digits. This was a vital step in creating a raw dataset that will be representative of human nature of speaking/writing Bengali.


বাংলাদেশ কৃষি ব্যাংক ও রাজশাহী কৃষি উন্নয়ন ব্যাংকের (রাকাব) শীর্ষ-২০ ঋণ খেলাপি গ্রাহকের কাছ থেকে আদায় হয়েছে মাত্র ২২ কোটি টাকা, যা লক্ষ্যমাত্রার ৭.০৯ শতাংশ। গত অর্থবছর ৩১০ কোটি টাকা আদায়ের লক্ষ্যমাত্রা দিয়েছিল কেন্দ্রীয় ব্যাংক।

অবশ্য শীর্ষ-২০ খেলাপির বাইরে অন্য খেলাপিদের থেকে আদায় সন্তোষজনক পর্যায়ে রয়েছে। তবে উচ্চমাত্রার খেলাপি ঋণ ও মূলধন ঘাটতি ব্যাংক দুটিকে লোকসান থেকে বের হতে দিচ্ছে না। এখন গুণগত মানের ঋণ বিতরণ ও আদায় জোরদারের মাধ্যমে পরিস্থিতির উন্নয়ন করতে বলা হয়েছে।

গত রবিবার বিশেষায়িত খাতের বাংলাদেশ কৃষি ও রাকাবের চেয়ারম্যান ও এমডিদের নিয়ে অনুষ্ঠিত বৈঠকে এসব আলোচনা হয়। বাংলাদেশ ব্যাংকের সভাকক্ষে অনুষ্ঠিত বৈঠকে সভাপতিত্ব করেন গভর্ন র ড. আতিউর রহমান। এ সময় ডেপুটি গভর্নর, নির্বাহী পরিচালকসহ বিভিন্ন পর্যায়ের কর্মকর্তারা উপস্থিত ছিলেন। মূলত ব্যাংক দুটির ২০১৪-১৫ অর্থবছরের সার্বিক সূচকের পর্যালোচনা ও উন্নয়নে ক রণীয় নির্ধারণ বিষয়ে এ বৈঠক ডাকা হয়।

Fig. 3.3: Before filtering unnecessary words

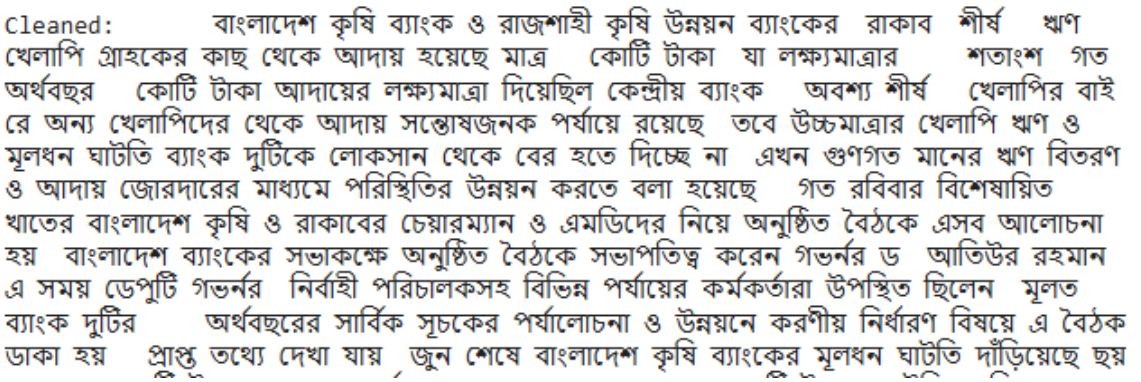

Cleaned: বাংলাদেশ কৃষি ব্যাংক ও রাজশাহী কৃষি উন্নয়ন ব্যাংকের রাকাব শীর্ষ ঋণ খেলাপি গ্রাহকের কাছ থেকে আদায় হয়েছে মাত্র কোটি টাকা যা লক্ষ্যমাত্রার শতাংশ গত অর্থবছর কোটি টাকা আদায়ের লক্ষ্যমাত্রা দিয়েছিল কেন্দ্রীয় ব্যাংক অবশ্য শীর্ষ খেলাপির বাই রে অন্য খেলাপিদের থেকে আদায় সন্তোষজনক পর্যায়ে রয়েছে তবে উচ্চমাত্রার খেলাপি ঋণ ও মূলধন ঘাটতি ব্যাংক দুটিকে লোকসান থেকে বের হতে দিচ্ছে না এখন গুণগত মানের ঋণ বিতরণ ও আদায় জোরদারের মাধ্যমে পরিস্থিতির উন্নয়ন করতে বলা হয়েছে গত রবিবার বিশেষায়িত খাতের বাংলাদেশ কৃষি ও রাকাবের চেয়ারম্যান ও এমডিদের নিয়ে অনুষ্ঠিত বৈঠকে এসব আলোচনা হয় বাংলাদেশ ব্যাংকের সভাকক্ষে অনুষ্ঠিত বৈঠকে সভাপতিত্ব করেন গভর্নর ড আতিউর রহমান এ সময় ডেপুটি গভর্নর নির্বাহী পরিচালকসহ বিভিন্ন পর্যায়ের কর্মকর্তারা উপস্থিত ছিলেন মূলত ব্যাংক দুটির অর্থবছরের সার্বিক সূচকের পর্যালোচনা ও উন্নয়নে করণীয় নির্ধারণ বিষয়ে এ বৈঠক ডাকা হয় প্রাপ্ত তথ্যে দেখা যায় জুন শেষে বাংলাদেশ কৃষি ব্যাংকের মূলধন ঘাটতি দাঁড়িয়েছে ছয়

Fig. 3.4: After filtering unnecessary words

- **Stop Word Removal:** Around 700 stop words of Bengali were used to remove words that occur very often but contribute nothing semantically. It is a step towards noise reduction and tend models to way more meaningful terms in the text.

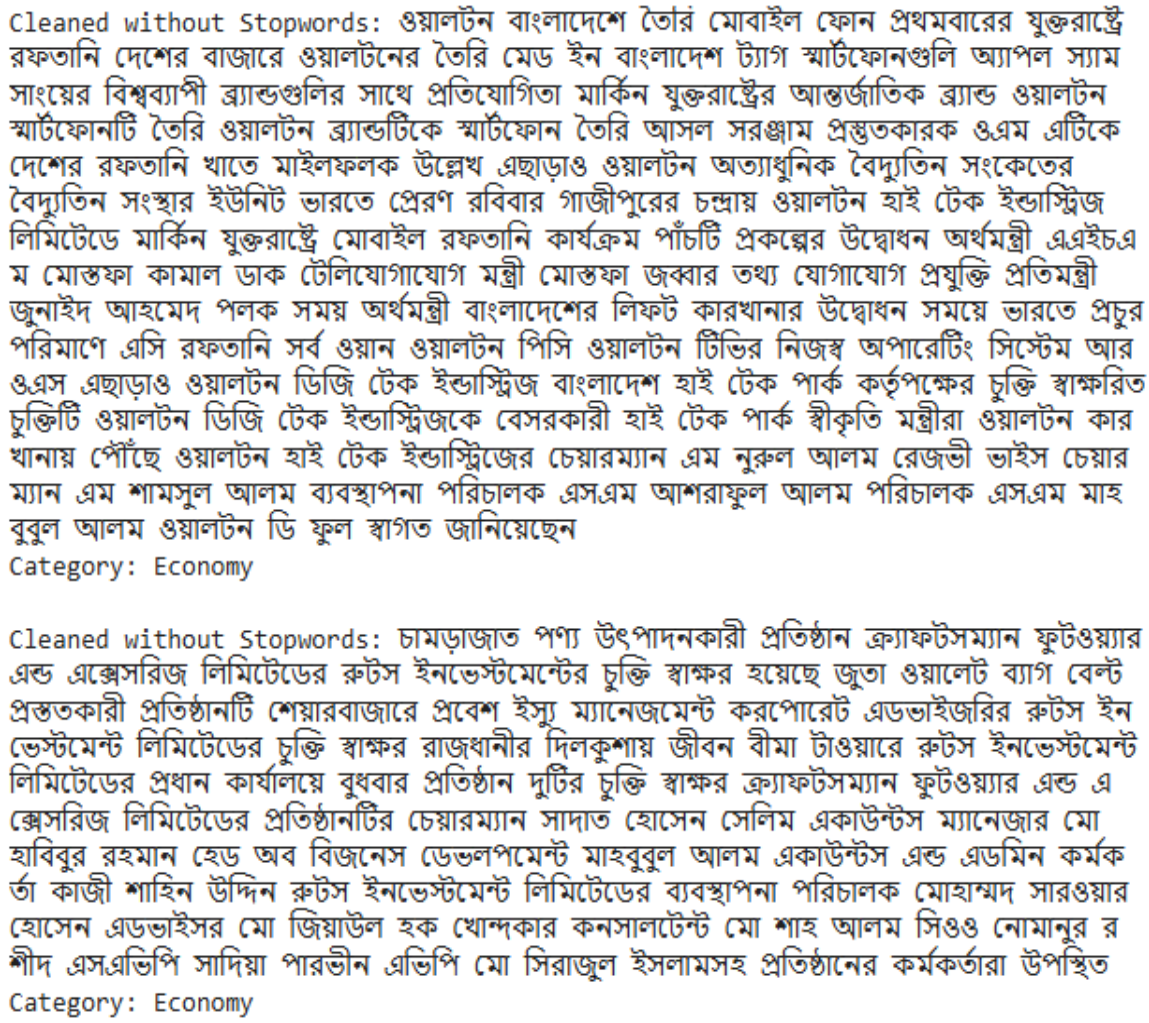


Fig. 3.5: Before and after removing the stop words

- **Short Article Removed:** We have removed 1785 small articles which have less than 20 words. After removing the articles which contains less than 20 words, we printed the total number of rows and length of each rows as shown in below.

	article	category	cleaned	length
0	সাভারের কবিরপুর বাণিজ্যিক এলাকার নিজস্ব ফ্যাক্...	Economy	সাভারের কবিরপুর বাণিজ্যিক এলাকার নিজস্ব ফ্যাক্...	157
1	\nরিজার্ভ চুরিচলতি বছরের সবচেয়ে চাঞ্চল্যকর ঘটন...	Economy	রিজার্ভ চুরিচলতি বছরের সবচেয়ে চাঞ্চল্যকর ঘটনা...	989
2	অর্থনৈতিক রিপোর্টার : এসএমই ফাউন্ডেশনের কনফারে...	Economy	অর্থনৈতিক রিপোর্টার এসএমই ফাউন্ডেশনের কনফারে...	130
3	ধীরে কমছে চট্টগ্রাম বন্দরের জট। টানা বর্ষণ, ব...	Economy	ধীরে কমছে চট্টগ্রাম বন্দরের জট টানা বর্ষণ ব...	371
4	\n\nমোবাইলফোনে কথা বলায় এবার খরচ বাড়ছে। বাজেটে...	Economy	মোবাইলফোনে কথা বলায় এবার খরচ বাড়ছে বাজেটে ম...	194
...	...	...	...	...
324962	\n\n পরবর্তী কর্মসূচি না দেয়া পর্যন্ত সারা...	politics	পরবর্তী কর্মসূচি না দেয়া পর্যন্ত সারাদে...	931
324963	\n\n প্রধানমন্ত্রী শেখ হাসিনার নামে ঢাকা বিশ্ব...	politics	প্রধানমন্ত্রী শেখ হাসিনার নামে ঢাকা বিশ্ববি...	276
324964	\n\n ভাই জিএম কাদেরকে দলের কো-চেয়ারম্যান করা...	politics	ভাই জিএম কাদেরকে দলের কো চেয়ারম্যান করা এ...	226
324965	\n\n তথ্যমন্ত্রী ও জাসদ সভাপতি হাসানুল হক ইন...	politics	তথ্যমন্ত্রী ও জাসদ সভাপতি হাসানুল হক ইনু ...	128
324966	\n\n বিএনপি নেতৃত্বাধীন ২০ দলীয় জোটের বৈঠক হবে...	politics	বিএনপি নেতৃত্বাধীন দলীয় জোটের বৈঠক হবে আজ...	43

324967 rows × 4 columns

Fig. 3.6: After removing short articles

- **Tokenization:** We tokenized every cleaned text entry, the process of breaking down a text into individual tokens (in our case, words) We have shown the summary of each document, word, unique word counts per category in the dataset after removing stop words and short articles. Here in the fig 3.7 the most frequent words with count are shown.

```
Class Name: International
Number of Documents: 40983
Number of Words: 6005625
Number of Unique Words: 213445
Most Frequent Words:

হয়েছে    57917
এক        43876
প্রেসিডেন্ট    25285
গত        23903
মার্কিন   22652
সময়      18632
কথা       17712
হয়ে      17562
দেশটির    17172
জানিয়েছে    16500
রয়েছে    16251
বিরুদ্ধে  16036
বছর       15948
ভারতের    15598
ট্রাম্প   15423
দিয়ে     14987
খবর       14967
হিসেবে    14358
পুলিশ     14138
জানান     13163
```

```
Class Name: Education
Number of Documents: 39644
Number of Words: 7211536
Number of Unique Words: 190150
Most Frequent Words:

খ         124856
গ         115879
ক         112551
ঘ         101944
কোনটি    30741
নিচের     24040
এক        22875
বিশ্ববিদ্যালয়ের    22867
হয়েছে    21260
অধ্যাপক   21005
উ         19145
সঠিক      18436
সালে      18275
বিভাগের   18141
যায়      17429
ড         17082
বিশ্ববিদ্যালয়ের    17017
টাকা      16252
ভর্তি     15707
পরীক্ষা   15304
```

```
Class Name: Entertainment
Number of Documents: 40615
Number of Words: 5149469
Number of Unique Words: 175098
Most Frequent Words:

হয়েছে    29085
অভিনয়    28001
এক        23695
কথা       19959
গান       17537
খান       16628
ছবি       16547
হয়ে      16198
চলচ্চিত্র   16039
সময়      14850
অভিনেত্রী    14705
বছর       14289
হিসেবে    13074
মুক্তি    12798
অভিনেতা    12453
ভালো      12433
ছবির      12318
যায়      12146
সিনেমা    11873
পরিচালক   11464
```

Fig. 3.7: Tokenization of each Documents

3.5 Data Insight Details

This dataset contains 4 attributes title, body, label, source while some variation and training purpose The text entry for and initial dataset is based on Potrika, However I am to decrease it and final version of the Potrika data consists of more than 3,20,000 data have been divided into 8 different category sports, national, international, education, science & technology, politics entertainment economic-business will be used in our experiment with various algorithms.

Table 3.1: Count of the articles in each category

Category	**Count**
ScienceTechnology	42083
International	40979
National	40928
Sports	40701
Entertainment	40539
Economy	40273
Politics	40104
Education	38533

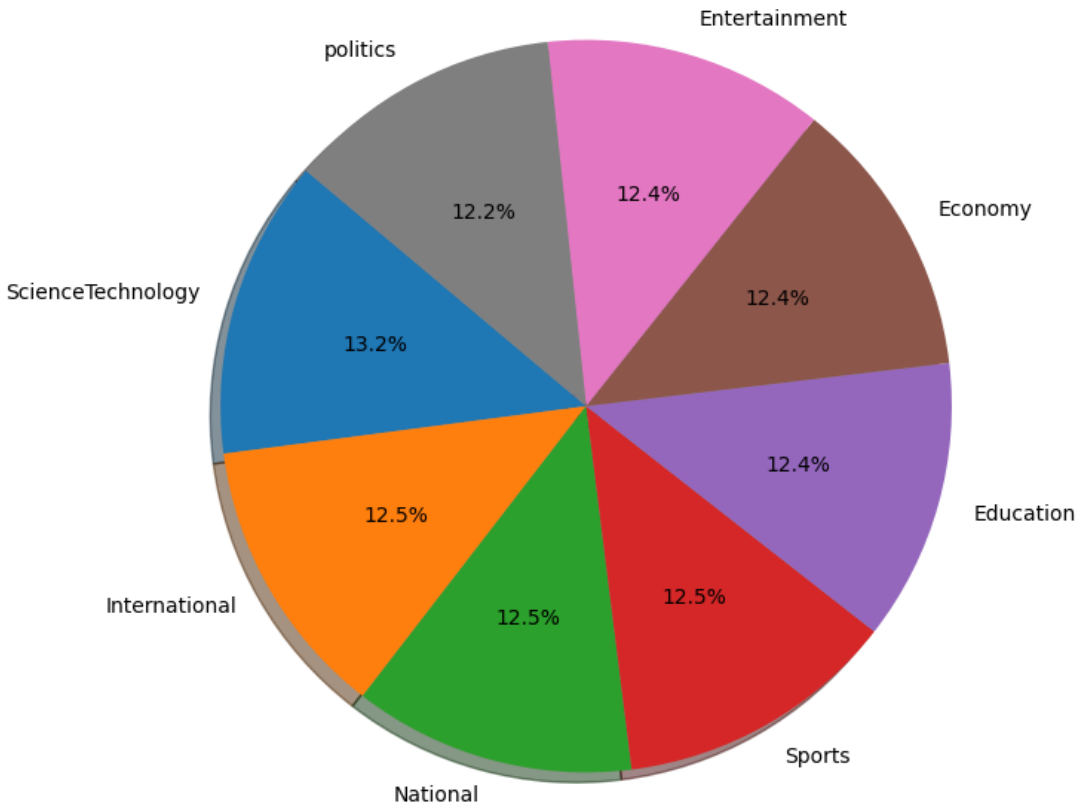


Fig. 3.8: Size of each category in pie-chart

3.6 Understanding of the Deep Learning Models

Recently, DL due to its imply can learn from the provided data; hence it has become one of the hot topics these days in ML, AI as well as DS and DA[14]. In this research we have used 5 different deep learning models building including CNN, LSTM, GRU, CNN-LSTM(Hybrid) and Bangla-BERT. Here I will explain all of the working principles of these methods.

3.6.1. Transformer: Bangla-BERT

The initial application of transformers was sequence transduction or in layman's terms: neural machine translation. So they are meant to work on any sequence-to-sequence task. That is why they are called "Transformers". The latest state-of-the-art NLP model, Transformers, is a gradual evolution of the encoder-decoder architecture. While the encoder-decoder model relies extensively on Recurrent Neural Networks (RNNs) to capture the time-dependent information in the input, Transformers have no such recurrence[15].

Bangla-BERT is a pre-trained model based on the BERT architecture, specifically adapted for the Bengali language[16]. It's a type of ELECTRA discriminator model that's been pre-trained with a special objective called Replaced Token Detection (RTD). This means it's really good at understanding the nuances of the Bengali language[17].Utilizing a transformer architecture, Bangla-BERT excels at capturing contextual information from text. The model comprises multiple layers of transformer blocks, each containing multi-head self-attention mechanisms and feed-forward neural networks. Pre-trained on an extensive Bengali text corpus, Bangla-BERT learns complex language representations, making it easy to fine-tune for specific downstream tasks with minimal additional training[18].

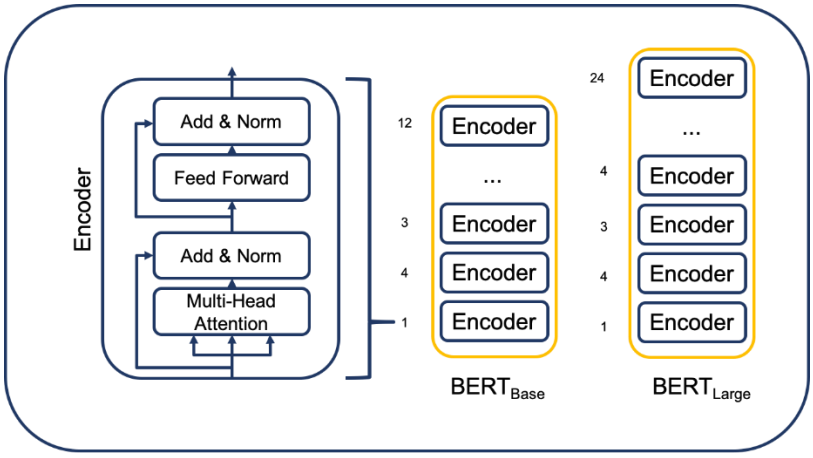


Fig. 3.9: Architecture of BERT Transformer

3.6.2. Convolution Neural Networks (CNN)

CNNs are layered artificial neural networks that can identify intricate features in data; for example, it finds features in image and text data. CNNs have primarily been applied to computer vision problems, including image classification, object detection, and image segmentation. But recently people used CNNs in text problems[19].

Language is sequential and high-dimensional in nature, meaning that when we work with text data (unstructured), we frequently have to deal with a huge vocabulary. However, before feeding this data into any CNNs, we need to preprocess it with techniques like tokenization, stemming/lemmatization and vectorization (like TF-IDF etc.)

A typical CNN architecture in NLP involves an embedding layer that converts words to dense vectors, convolutional layers which use filters over the embedded text, pooling layers (max or average are common choices) which down-sample the representation, fully connected layers which interpret those features and a final output layer for classification. Together they help us understand context when reading texts. CNNs are trained on labelled data, which means that for each text denominator we have categories. Loss function, which is a measure of the difference between predicted labels and their respective classes, can be minimized using backpropagation and gradient descent algorithms[20]. Through this process, the model adjusts its weights in an iterative fashion.

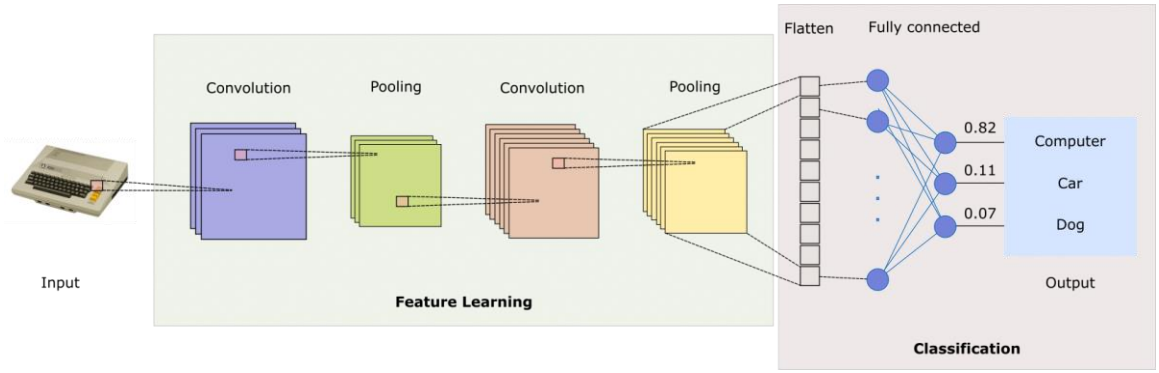


Fig. 3.10: Architecture of CNN

3.6.3 Recurrent Neural Networks (RNN)

Recurrent Neural Networks: A specialized type of artificial neural network that has been designed to perform well on sequential data. Its commonly used in natural language processing, such as language translation, Speech Recognition, Sentiment Analysis, Natural Language Generation and Text Summarization. Recurrent working function: RNN memory introduces a loop or cycle that has some built-in structure which allows the model to remember information over time unlike neural network feedforward. This makes them unlike feedforward neural networks.

Recurrent Neural Networks (RNNs) are a highly valuable class of artificial neural networks in the field of Natural Language Processing (NLP). RNNs play an important role mainly in text classification tasks[21]. Another point that sets RNNs apart from standard feedforward networks, and as a universal approximation in general, is that they can fit sequential dependencies of data, which makes them suitable for processing sequences such as language. In NLP text classification, RNNs tend to do well in assessing the contextual relationships between words as they are able to predict patterns and semantics that are vital for accurately classifying textual information. RNNs are fundamental for building complex models to classify documents, detect spam and analyze sentiment due to their versatility.

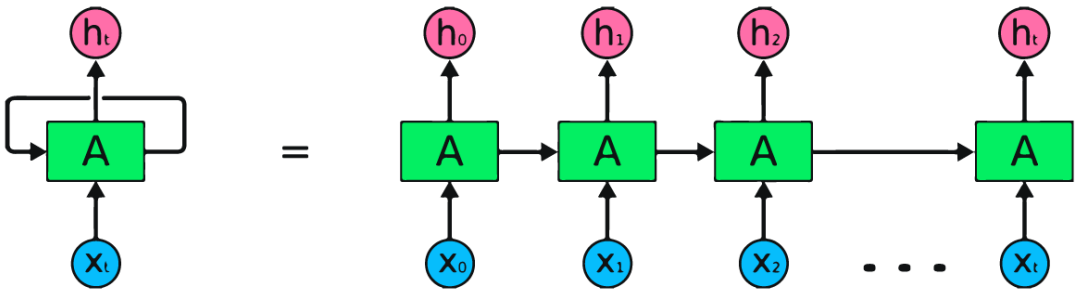


Fig. 3.11: Architecture of RNN

3.6.4. Gated Recurrent Unit (GRU)

Gated Recurrent Units or just GRUs, like LSTMs are a type of RNN that is also capable of learning long-term dependencies and dealing with sequential data but have a simplified architecture. A GRU has two main gates: the update gate and the reset gate, which it uses to receive input sequences (for example – word embeddings)[22]. The update gate decides the amount of information to pass along from the past hidden state to the current time step, while the reset gate determines how much past memory needs to be "forgotten". These gates enable GRUs to dynamically control the passing of information, maintaining relevant context while overcoming the vanishing gradient problem. GRU is the same as RNN but in GRU, we use gates to control the flow of information into and out of the hidden state, which helps with retaining information from previous time steps. In text classification, like mentioned above GRUs processes words one by one and updates its hidden states at each step. The final hidden state (or a pooled version of it) is then processed in an additional dense layer with softmax activation for class label prediction (e.g. Bangla Newspaper Classification) after the entire sequence has been processed and passed through the network. GRUs are also widely used for sequential data problems like LSTM, but they come with a simple

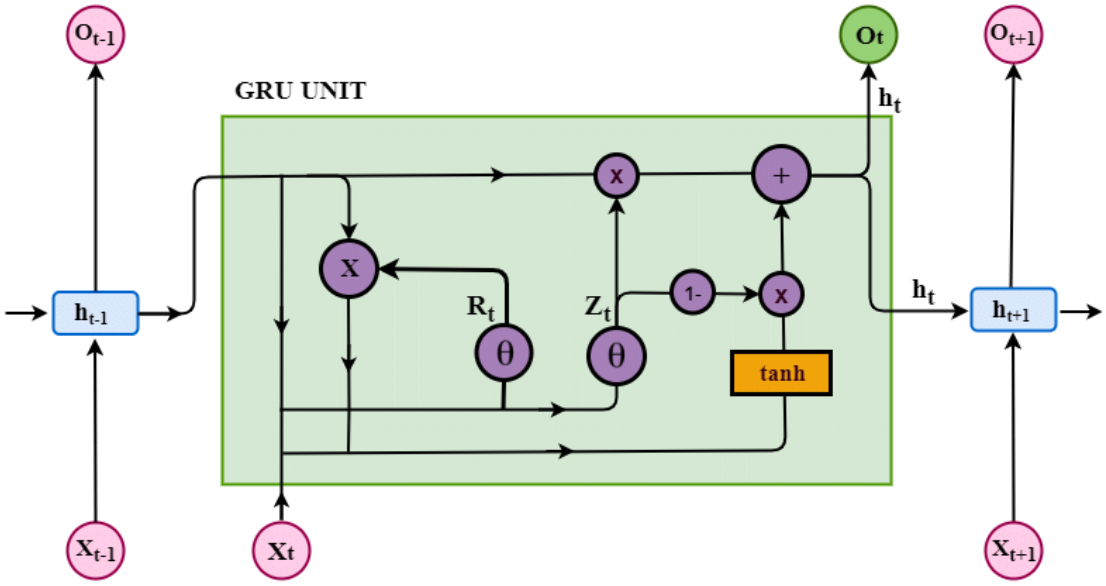


Fig. 3.12: Architecture of GRU

3.6.5 Long Short-Term Memory (LSTM)

LSTM or Long Short-Term Memory is a kind of recurrent neural network which is more memory-friendly than vanilla RNNs. LSTMs perform fairly better having a good hold over memorizing certain patterns[23].

LSTMs handle sequential data, like text in classifications where information about context matters (like if a series of words that A and B create have no relation to C or D). To overcome this limitation of traditional Recurrent Neural Networks (RNNs), which struggle to capture long-range dependencies due to the vanishing gradient problem, LSTMs are introduced. LSTM operates by feeding sequences of inputs (in this case, word embeddings) through a number of gates (forget, input and output gate) that determines what information goes in and out of the cell state. This allows the LSTM to "retain" valuable information across long distances in the sequence, whilst "losing" redundant details. For a standard text classification problem, LSTM reads through the words of a sequence one by one, carrying forward an updated hidden state. At the end of processing the entire sequence, we feed the final hidden state or its aggregate states through a fully connected layer followed by softmax activation to predict one class label, e.g. sentiment or topic category. LSTMs suit exceptionally well for tasks such as sentiment classification, document categorization, or any task that contains the sequence order of words and the contextual meaning of such in your text.

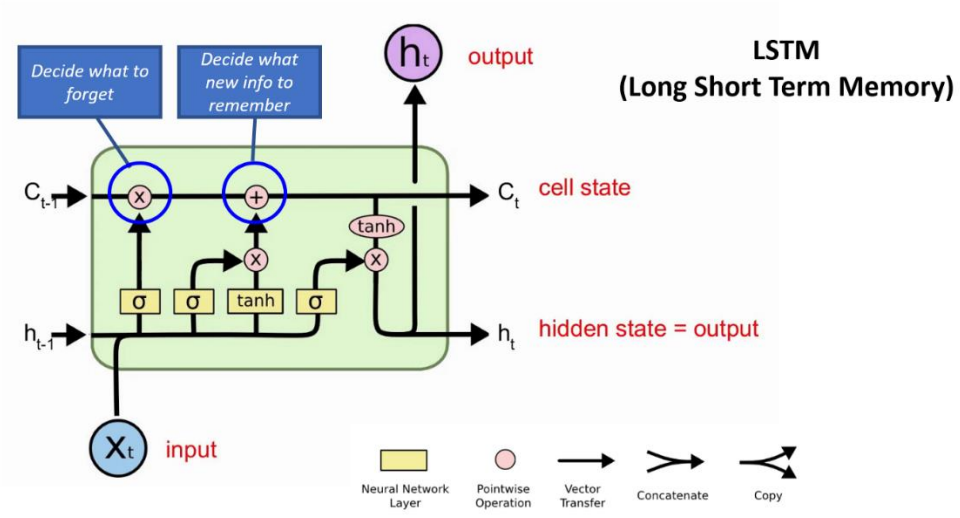


Fig. 3.13: Architecture of LSTM

3.6.6. Hybrid (CNN+LSTM)

Bengali text classification using a hybrid model (Convolutional Neural Networks- CNN + Recurrent Neural Networks-RNN) In this architecture, the CNN first captures local features and patterns in a text for instance representing n-grams or character-level features by applying convolutional filters to the input text[24]. The CNN is efficient because it is able to capture spatial hierarchies and also local dependencies present within the text, that makes them very helpful in detecting semantic features such as some important phrases or keywords. The extracted features from the preceding layer are then fed into the RNN layer (usually LSTM or GRU) to account for sequential dependencies and contextual information across longer distances in the text which is essential for capturing broader context of a sentence/doc. In the Bengali text classification scenario, this combined model form is very useful because of the morphological richness and unique script in Bengali where a classifier needs local feature extraction (in our case CNN) to identify some identifying features while it also requires long-term contextual understanding (in our case RNN) in order to classify articles into different categories such as politics, sports, or entertainment. Since CNN will be able to learn local patterns and RNN learns global context making this a strong approach for Bengali text classification tasks. The resulting feature is usually fed to a fully connected layer in order to estimate the correct class label.

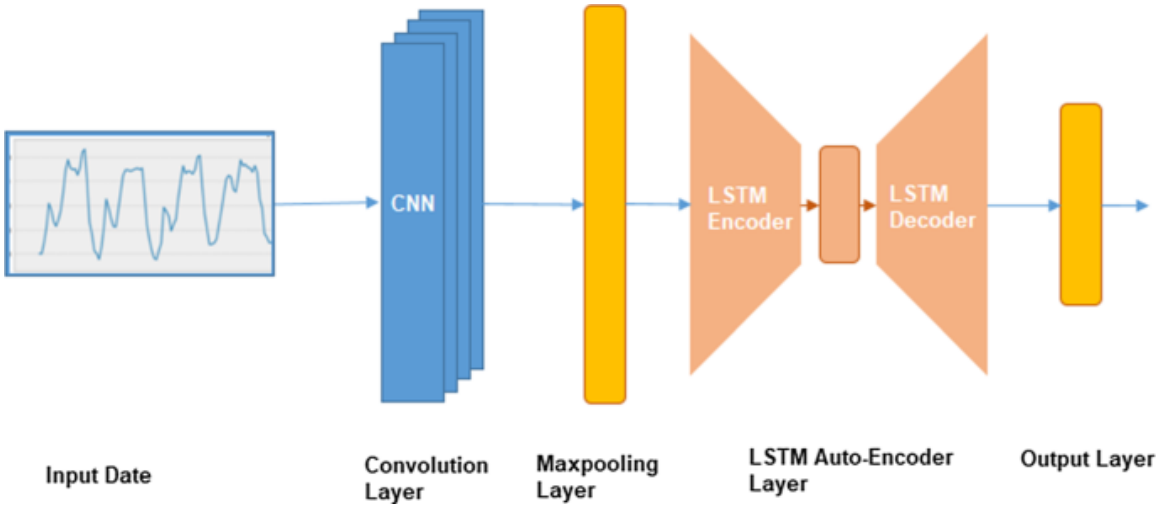


Fig. 3.14: Architecture of Hybrid Model (CNN+LSTM)

3.7 Training Model

In this study, deep learning models like LSTM, GRU, RNN, CNN, and Hybrid models were used, with 20% of the dataset reserved for testing. LSTM and GRU models utilized 4 neural layers, with tokenized inputs and 20% dropout to reduce overfitting. Bangla-BERT, a transformer-based model, was trained with a 64 batch size and 5 epochs, processing tokenized words through an embedding layer.

3.8 Implementation Requirements

After an in-depth review of the relevant statistical and theoretical concepts, a list of essential prerequisites for this text classification project was identified. The requirements include:

Hardware/Software Requirements

- Operating System: Windows 7 or newer
- Hard Disk: Minimum of 30 GB
- RAM: Minimum of 12 GB

Development Tools

- Python Environment
- Jupyter Notebook
- Google Colab

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

The performance metric value obtained when the model is trained and evaluated on the training dataset is called as training accuracy. It depicts the amount of times the model learned from the data it was trained upon. After training has finished, the model is tested against the testing dataset that it has not seen previously. How well this new data performs is called test accuracy, which gives us insight into how well the model generalizes to unseen data. We then created a plot with respect to this model that shows training and test accuracy over time.

Our dataset consists of more than 324000 data, where each category contains 30000-40000 article. But we have taken 5000 random sample from each of the 8 categories for the experimental analysis. Then within this 40000 samples article we have run our deep learning models.

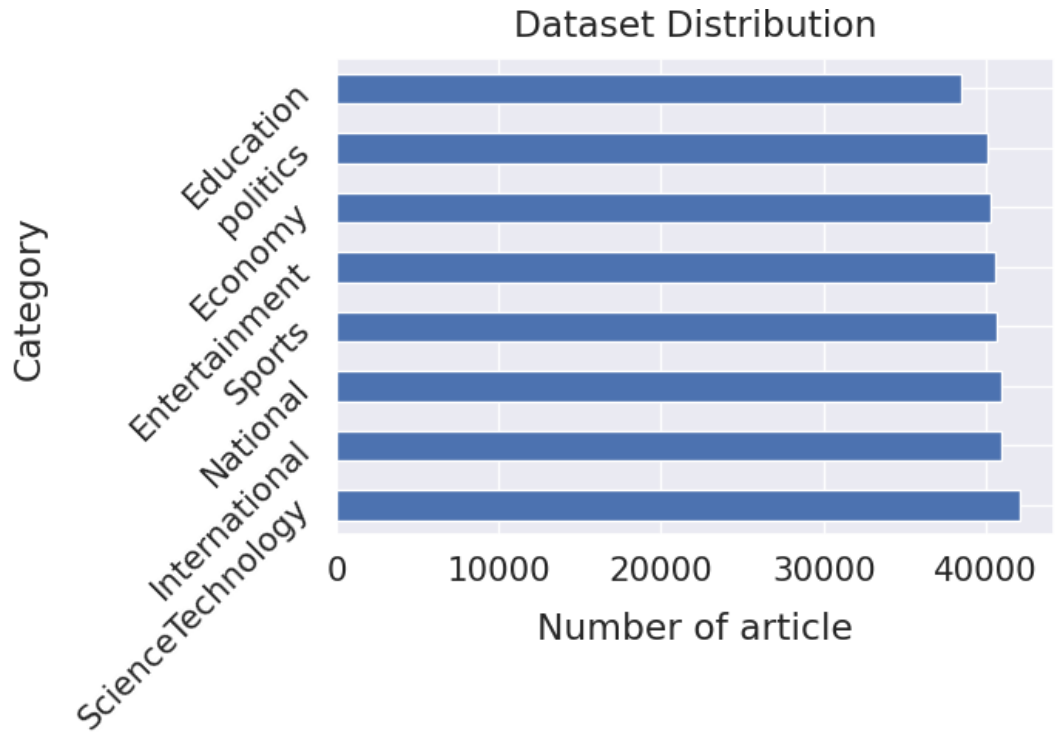


Fig. 4.1: Dataset Distribution of each category

	cleanedwithoutstopwords	category
0	অর্থনৈতিক রিপোর্টার মিডল্যান্ড ব্যাংক লিমিটেড ...	Economy
1	বাণিজ্য ভারসাম্যে ঘাটতি রেমিট্যান্স প্রবাহ হ্র...	Economy
2	সিরাজগঞ্জের বেলকুচিতে গত সোমবার মাইচয়েস মাইওয়া...	Economy
3	ঈদুল আজহার ন্যূনতম দিনের বাংলাদেশ ভারতসহ দেশেই...	Economy
4	প্রধানমন্ত্রীর কার্যালয়ের সচিব মো তোফাজ্জল হোস...	Economy
...	...	...
39995	যুগ্ম সচিব পদে পদোন্নতির ঘটনাকে সম্পূর্ণভাবে স...	politics
39996	জাতীয় সংসদকে পুতুল নাচের নাট্যশালা বলায় ট্রান্...	politics
39997	সাবেক মন্ত্রী ওয়ার্কার্স পার্টির সভাপতি রাশেদ...	politics
39998	বিএনপির ভারপ্রাপ্ত চেয়ারম্যান তারেক রহমান সংস...	politics
39999	একদলীয় শাসন কায়েমের ধারাবাহিকতায় শফিক রেহমানকে...	politics

40000 rows × 2 columns

Fig. 4.2: Sample dataset to run the models

4.2 Evolution Methods

A confusion matrix summarizes how well a machine learning model performs over a set of test data. It shows instances that are correctly and incorrectly predicted by the model. The most popular usage of it is to evaluate classifiers that assign a label (finite number of values) to each input.

- True Positive (TP): The model predicted positive and it was actually positive.
- True Negative (TN): The actual outcome was negative and the model predicted negative.
- False Positive (FP): The model made a positive prediction (the actual outcome was negative). Also known as a Type I error.

- FN (False Negative) − Model predicted negative but actual belongs to the positive class. This is also called a Type II error.

Metrics based on Confusion Matrix Data:

- **Accuracy**

Accuracy is used to measure the performance of the model. It is the ratio of Total correct instances to the total instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**

Precision is a measure of how accurate a model's positive predictions are. It is defined as the ratio of true positive predictions to the total number of positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**

Recall measures the effectiveness of a classification model in identifying all relevant instances from a dataset. It is the ratio of the number of true positive (TP) instances to the sum of true positive and false negative (FN) instances.

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score**

F1-score is used to evaluate the overall performance of a classification model. It is the harmonic mean of precision and recall,

$$F1 - Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

4.3 Experimental Results & Analysis

4.3.1 Bangla-BERT

With an accuracy of 92.00%, it outperforms other models by a significant margin, making it the best choice for tasks that require a high level of accuracy and contextual understanding in Bangla. The confusion matrix, as shown in Figure 4.3.1. This figure provides a detailed view of the model's performance.

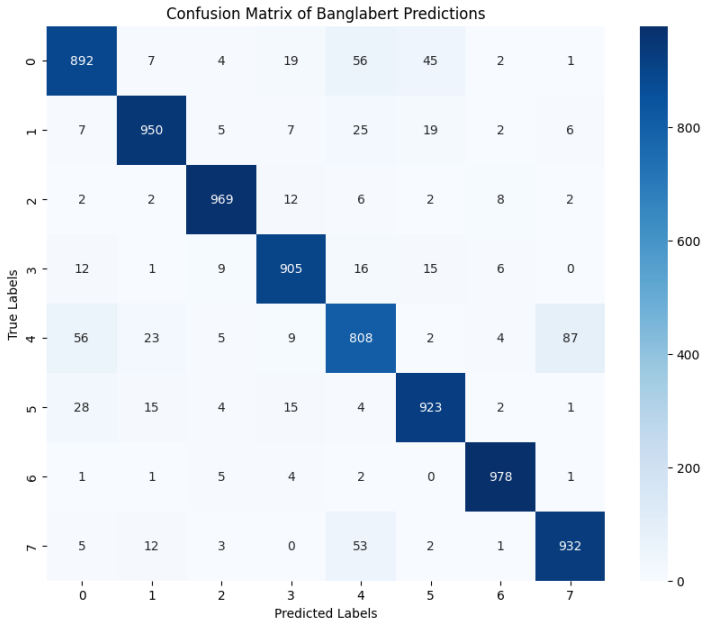


Fig. 4.3.1: Confusion Matrix of Bangla-BERT

Classification Report of Banglabert Prediction:

	precision	recall	f1-score	support
Economy	0.89	0.87	0.88	1026
Education	0.94	0.93	0.94	1021
Entertainment	0.97	0.97	0.97	1003
International	0.93	0.94	0.94	964
National	0.83	0.81	0.82	994
ScienceTechnology	0.92	0.93	0.92	992
Sports	0.98	0.99	0.98	992
politics	0.90	0.92	0.91	1008
accuracy			0.92	8000
macro avg	0.92	0.92	0.92	8000
weighted avg	0.92	0.92	0.92	8000

Fig. 4.3.2: Classification Report of Bangla-BERT

4.3.2 CNN

With an accuracy of 88.54%, CNN performs well, utilizing convolutional layers to identify important n-grams or local features, although it does not capture long-term dependencies as effectively as BERT. We run the CNN model using 64 batch size with 5 epochs. The confusion matrix, as shown in Figure 4.3.3 This figure provides a detailed view of the model's performance.

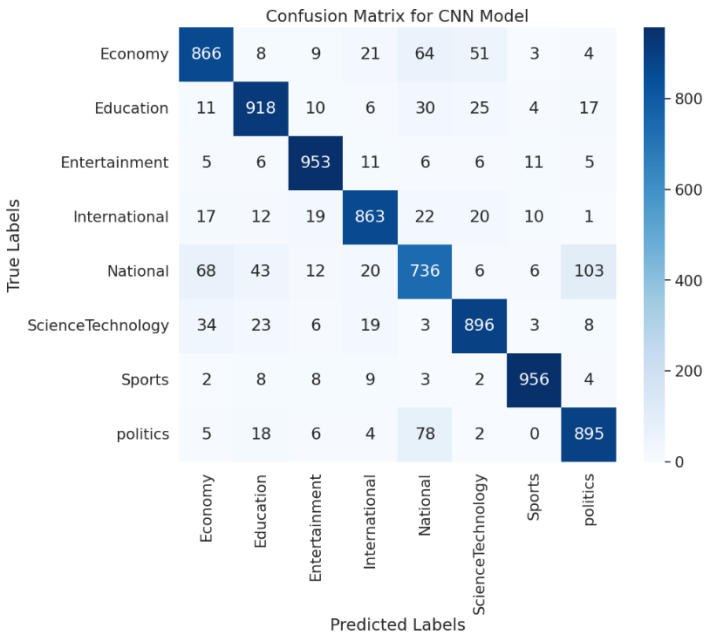


Fig 4.3.3: Confusion Matrix of CNN

Classification Report for CNN Model:

	precision	recall	f1-score	support
Economy	0.86	0.84	0.85	1026
Education	0.89	0.90	0.89	1021
Entertainment	0.93	0.95	0.94	1003
International	0.91	0.90	0.90	964
National	0.78	0.74	0.76	994
ScienceTechnology	0.89	0.90	0.90	992
Sports	0.96	0.96	0.96	992
politics	0.86	0.89	0.88	1008
accuracy			0.89	8000
macro avg	0.88	0.89	0.89	8000
weighted avg	0.88	0.89	0.88	8000

Fig 4.3.4: Classification Report of CNN

4.3.3 Gated Recurrent Unit (GRU)

We run the GRU model using 64 batch size and 5 epochs. The GRU model achieves 86.69% accuracy, indicating its ability to capture sequential dependencies, but it falls short of CNN and Bangla-BERT in overall performance. The confusion matrix, as shown in Figure 4.3.5. This figure provides a detailed view of the model's performance.

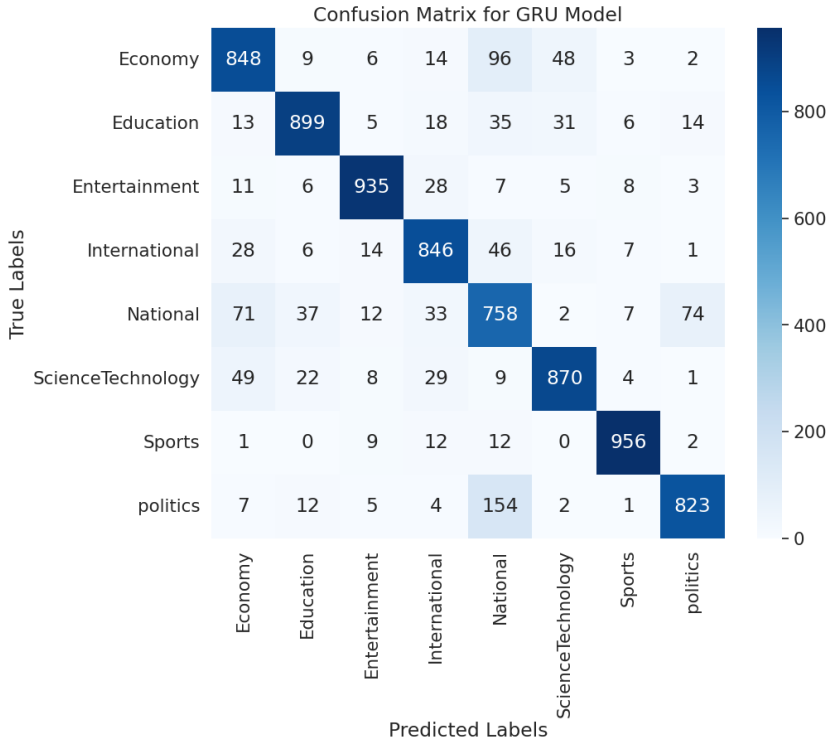


Fig 4.3.5: Confusion Matrix of GRU

Classification Report of GRU Model:

	precision	recall	f1-score	support
Economy	0.82	0.83	0.83	1026
Education	0.91	0.88	0.89	1021
Entertainment	0.94	0.93	0.94	1003
International	0.86	0.88	0.87	964
National	0.68	0.76	0.72	994
ScienceTechnology	0.89	0.88	0.89	992
Sports	0.96	0.96	0.96	992
politics	0.89	0.82	0.85	1008
accuracy			0.87	8000
macro avg	0.87	0.87	0.87	8000
weighted avg	0.87	0.87	0.87	8000

Fig 4.3.6: Classification Report of GRU

4.3.4 LSTM

We run LSTM model using 64 batch size and 5 epochs. With an accuracy of 84.25%, LSTM is the least accurate among these models, which may indicate limitations in processing Bangla text compared to other models that can capture local patterns and contextual information more effectively. The confusion matrix, as shown in Figure 4.3.7 This figure provides a detailed view of the model's performance.

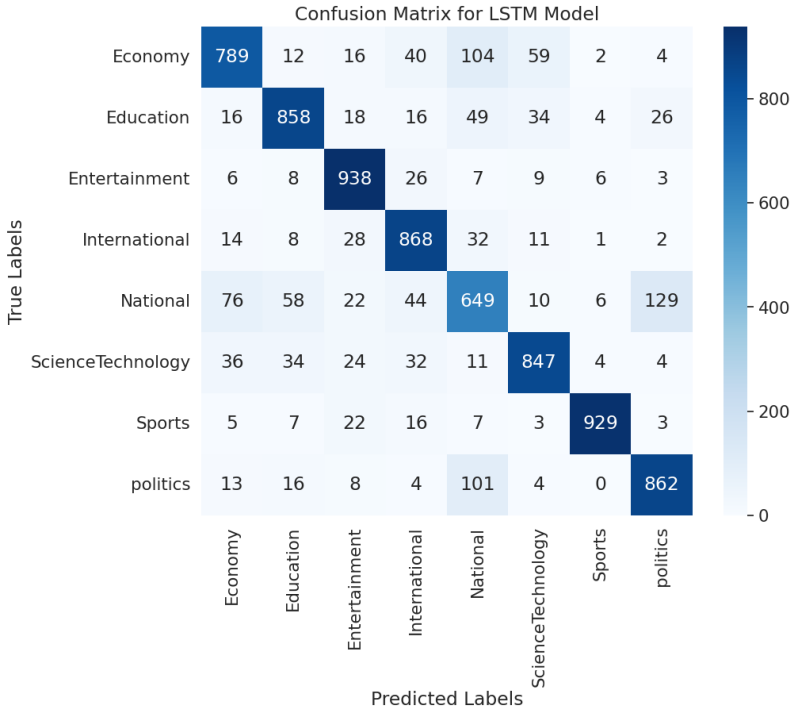


Fig 4.3.7: Confusion Matrix of LSTM

Classification Report of LSTM Model:

	precision	recall	f1-score	support
Economy	0.83	0.77	0.80	1026
Education	0.86	0.84	0.85	1021
Entertainment	0.87	0.94	0.90	1003
International	0.83	0.90	0.86	964
National	0.68	0.65	0.66	994
ScienceTechnology	0.87	0.85	0.86	992
Sports	0.98	0.94	0.96	992
politics	0.83	0.86	0.84	1008
accuracy			0.84	8000
macro avg	0.84	0.84	0.84	8000
weighted avg	0.84	0.84	0.84	8000

Fig 4.3.8: Classification Report of LSTM

4.3.5 Hybrid (LSTM-CNN)

With an accuracy of 87.88%, the hybrid model strikes a balance between local and sequential dependencies but doesn't perform as well as the CNN or Bangla-BERT model. The confusion matrix, as shown in Figure 4.3.9. This figure provides a detailed view of the model's performance.

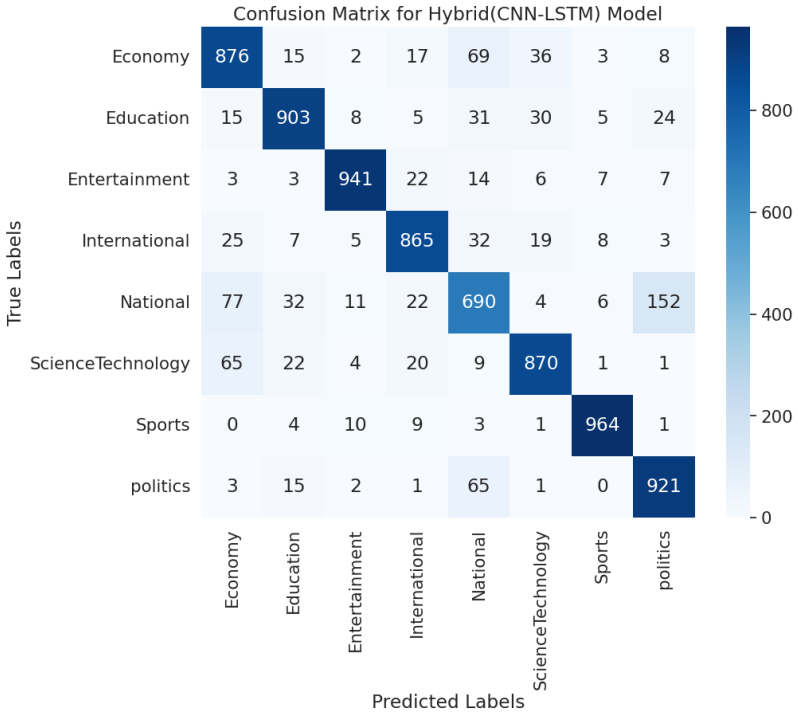


Fig 4.3.9: Confusion Matrix of Hyrid Model

Classification Report for Hybrid(CNN-LSTM) Model:

	precision	recall	f1-score	support
Economy	0.82	0.85	0.84	1026
Education	0.90	0.88	0.89	1021
Entertainment	0.96	0.94	0.95	1003
International	0.90	0.90	0.90	964
National	0.76	0.69	0.72	994
ScienceTechnology	0.90	0.88	0.89	992
Sports	0.97	0.97	0.97	992
politics	0.82	0.91	0.87	1008
accuracy			0.88	8000
macro avg	0.88	0.88	0.88	8000
weighted avg	0.88	0.88	0.88	8000

Fig 4.3.10: Classification Report of Hybrid Model

4.4 Comparison of the Model

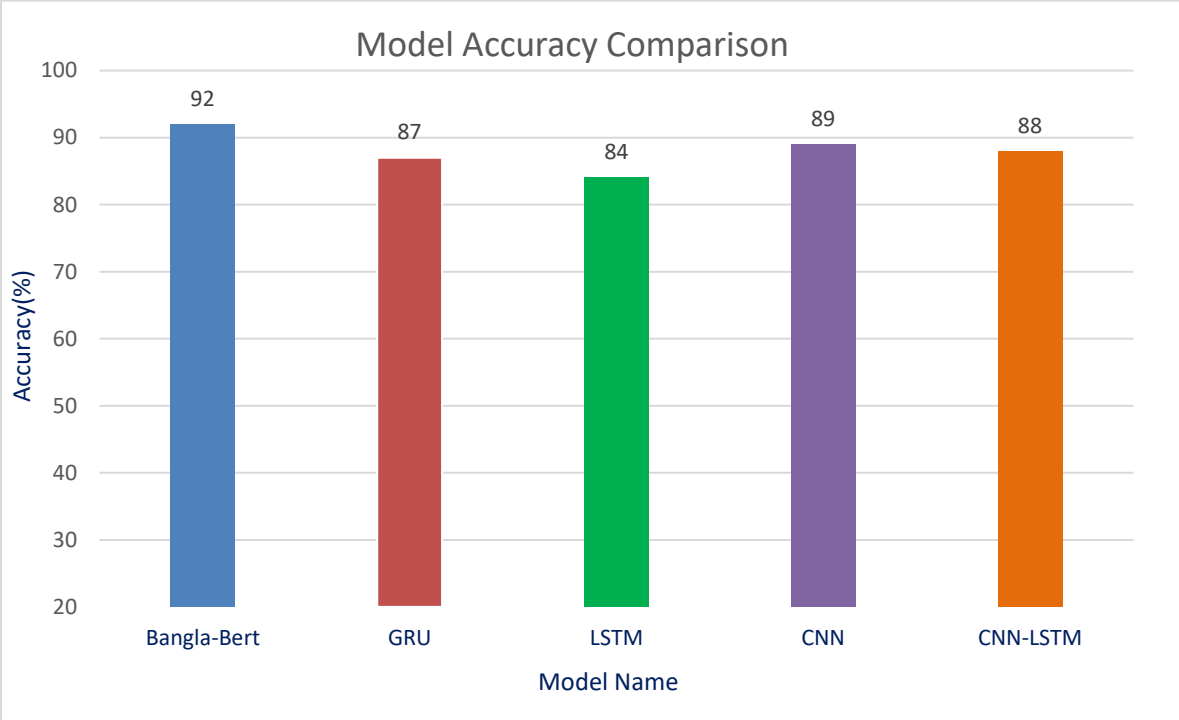


Fig 4.4: Model accuracy comparison

Table 4.1: Finding the best result among the Results of Deep Learning Models

Model	Test Accuracy (%)	Loss (%)	Precision (%)	Recall (%)	F1 Score (%)
Bangla-BERT	92.00	0.36	92.00	92.00	92.00
CNN	88.54	0.49	88.00	89.00	88.00
Hybrid	87.88	0.56	88.00	88.00	88.00
GRU	86.69	0.70	87.00	87.00	87.00
LSTM	84.25	0.73	84.00	84.00	84.00

Here’s a comparison of the models based on their accuracy levels for Bangla article classification:

4.4.1 Bangla-BERT (Accuracy: 92.00%)

- **Analysis**: Bangla-BERT has the highest test accuracy, precision, recall, and F1 score among the models. This indicates that Bangla-BERT performs best in terms of both accuracy and consistency across all evaluation metrics, making it the top choice for this classification task.

4.4.1 CNN (Convolutional Neural Network) (Accuracy: 88.54%)

- **Analysis**: CNN has relatively high accuracy and low loss, ranking just below Bangla-BERT. Its slightly higher recall indicates it may be better at identifying all relevant articles compared to GRU and LSTM. CNN offers a good balance between performance and computational efficiency.

4.4.3 Hybrid Model (Accuracy: 87.88%)

- **Analysis**: The hybrid model combines features from multiple architectures, achieving moderate accuracy and loss. Its precision, recall, and F1 scores are consistent at 88%, suggesting stable performance slightly below CNN but better than GRU and LSTM.

4.4.4 GRU (Gated Recurrent Unit) (Accuracy: 86.69%)

- **Analysis**: GRU performs moderately well, with slightly lower accuracy and higher loss compared to Bangla-BERT. Its precision, recall, and F1 score are consistent at 87%, indicating stable but less optimal performance. It’s a reasonable choice if computational resources are limited, though it doesn’t match Bangla-BERT ’s performance.

4.4.5 LSTM (Long Short-Term Memory) (Accuracy: 84.25%)

- **Analysis**: LSTM has the lowest accuracy and highest loss among the models, indicating it struggles more with this classification task. Its uniform precision, recall, and F1 score suggest that it performs consistently but is less effective than other models in capturing relevant patterns in the data.

Bangla-BERT leads with the highest accuracy, leveraging its contextual understanding capabilities. CNN follows with 88.54%, making it a suitable choice for slightly less computationally intensive tasks. Hybrid and GRU models provide balanced alternatives, while LSTM, with the lowest accuracy, may be less ideal for complex classification tasks in Bangla.

4.5 Descriptive Analysis

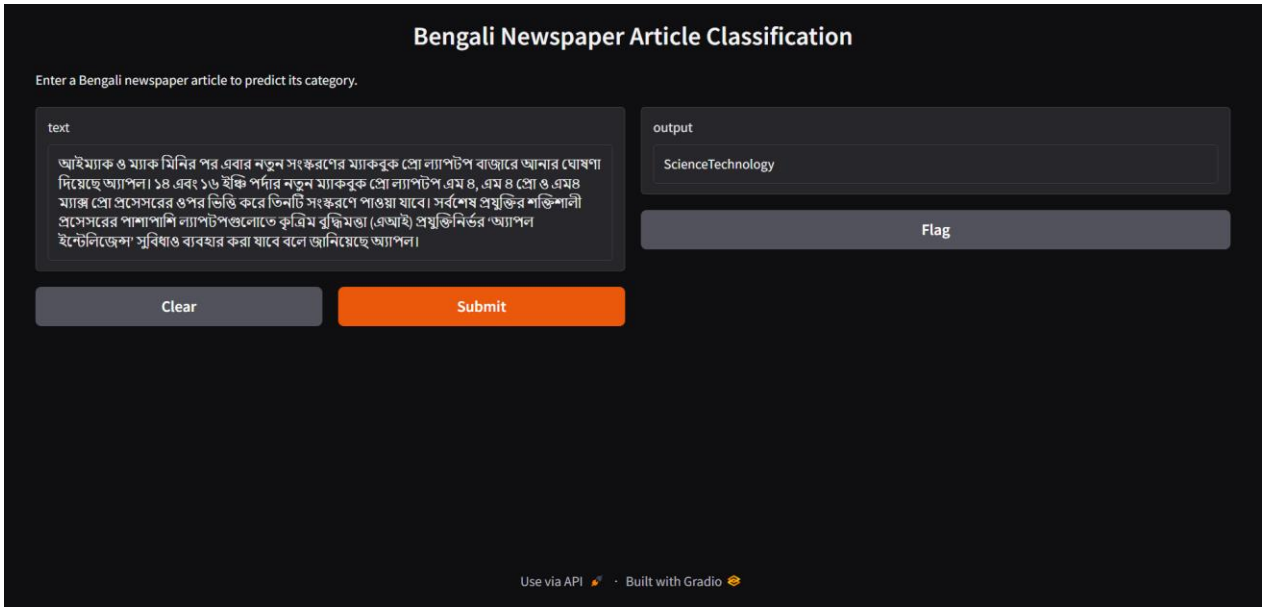


Fig 4.5: UI for predicting newspaper article headline

UI Architecture: Gradio, an open-source Python web UI library, helps us to bridge that gap between LLMs and non-technical end users. This enables us to create rapid prototypes for our DL projects making it easier to deploy them to a wider audience. This UI is designed

using the Gradio library in Python, hosted on Google Colab, to classify Bangla newspaper articles into categories. This Gradio-based UI is simple, user-friendly, and functional for testing and deploying Bangla article classification models in an accessible way.

Functionality and User Flow

- The user enters a Bangla newspaper article in the text input box.
- The user clicks Submit, and the model processes the article.
- The predicted category appears in the output box, as seen with the example category *"ScienceTechnology"*.
- The user can use Clear to reset the input or Flag to provide feedback on an incorrect classification.

4.6 Discussion

Best Performer: Bangla-BERT achieves the highest results across all metrics, making it the ideal choice for Bangla newspaper article classification.

Alternative Choices: CNN and Hybrid models provide a balance of accuracy and computational efficiency. CNN's slightly better recall may make it suitable for cases where identifying all relevant articles is critical.

Lower Performers: GRU and LSTM show lower accuracy and higher loss, indicating they might not be as effective for this task.

In summary, Bangla-BERT stands out as the best-performing model, excelling in accuracy, precision, recall, and F1 score. Its architecture is particularly well-suited to handling the nuances of Bangla newspaper articles. CNN and the Hybrid model offer respectable performance and could be considered as alternatives if computational resources are limited. GRU and LSTM, while competent, exhibit lower accuracy and higher loss, making them less effective for this specific classification task.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on society

By increasing media literacy, facilitating more effective communication, and expanding information access for Bengali-speaking communities, my research on Bengali newspaper classification has the potential to have a big social impact. Through the creation of sophisticated AI models specifically suited to the Bengali language, my work equips millions of Bengali speakers with the means to sift through enormous volumes of news content, facilitating their ability to remain informed on pertinent subjects like politics, health, education, sports, entertainment, international, national, science & technology and economy. By automating news classification, saving time, and freeing up news organizations to concentrate on content development, this not only helps to improve public awareness but also encourages more effective journalism. Additionally, my research can support varied viewpoints, raise awareness of significant social concerns, and promote social change by organizing and making news more accessible. Essentially, my work promotes an inclusive digital environment that benefits people and society at large by bridging the gap between language diversity and technology progress.

5.2 Impact on the environment

By diminishing the ecological footprint of conventional media methods, my paper on deep learning techniques can have a positive environmental impact. My work can assist move the emphasis from the creation of physical newspapers to digital media by automating the classification and organizing of news material. This will reduce the amount of paper waste, ink used, and energy used for printing and dissemination. Furthermore, effective digital news classification can simplify information retrieval, cutting down on pointless server load and internet traffic, which can help data centers use less energy. My study supports eco-friendly media practices and encourages the wider adoption of digital platforms over resource-intensive, paper-based systems by advocating for a more sustainable approach to information distribution.

5.3 Ethical Aspects

The ethical elements of AI and media can be greatly improved by multiclass classification of newspaper articles. My work encourages inclusivity by creating cutting-edge AI models especially for the Bengali language, guaranteeing that underprivileged populations and non-English speakers can benefit technological developments. Furthermore, by enabling more precise, consistent, and transparent categorization of news material, my study can aid in the fight against problems like media bias and disinformation. This can promote a more moral media environment where news is appropriately labeled, facilitating the public's ability to recognize reliable information. Furthermore, my research helps prevent biases that are frequently present in language models trained exclusively on English data by creating a model that is linguistically and culturally responsive, thereby increasing fairness and lowering the possibility of marginalizing marginalized voices. In the end, my work helps ensure that AI is used responsibly, serving the public interest while respecting media and technological ethics.

5.4 Sustainability Plan

My research's sustainability plan prioritizes ethical use, adaptability, and long-term effect. Regular updates will be done using fresh data to take into consideration changing linguistic patterns and new subjects in the Bengali media ecosystem in order to preserve and enhance the model over time. Making the model open-source would enable contributions from the international scientific community, guaranteeing ongoing enhancements and broader use. The classification system will be included into actual media processes through partnerships with news organizations, encouraging useful use and guaranteeing the tool's continued relevance in the media sector. A crucial element will be ethical supervision, with continuous audits to rectify any prejudices and guarantee openness in the classification of news. Moreover, the project's scalability will be a top goal in order to widen its footprint by allowing adaption across distinct languages and geographic regions. These initiatives will help numerous organizations advance this research, and support the appropriate expansion of AI in the media sphere.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of the Study

The findings indicate that transformer-based models, especially Bangla-BERT stands out as the best model for Bangla newspaper article classification, proving the effectiveness of transformer-based approaches over traditional architectures like LSTM, GRU, and CNN in this context. This study suggests that Bangla-BERT is the most suitable model for accurate and efficient Bangla text classification, setting a benchmark for future research in Bengali NLP applications.

6.2 Conclusions

The findings indicate that transformer-based models, specifically Bangla-BERT, outperform traditional deep learning models with 92% accuracy in Bengali news classification tasks. This underscores the importance of utilizing language-specific pre-trained models for enhanced accuracy. Along with this, other deep learning methods such as LSTM, CNN, GRU also performs well. Future research should explore ensemble methods combining these models to further improve classification performance.

6.3 Implication for Further Study

In this study, eight news categories were used across all models. Our future goal is to develop a more advanced neural network and expand the dataset with additional news categories. By incorporating more classification categories, we aim to improve the model's ability to build a more comprehensive classification framework. Additionally, we plan to apply ensemble techniques, combining the top-performing models to potentially achieve even higher accuracy, as ensemble may provide the best overall performance.

REFERENCES

[1] D. Tribune, “Bangla ranked at 7th among 100 most spoken languages worldwide,” Dhaka Tribune, Feb. 17, 2020. https://www.dhakatribune.com/world/201648/bangla-ranked-at-7th-among-100-most-spoken (accessed Nov. 12, 2024).

[2] Wikipedia, “Bengali language movement,” Wikipedia, Feb. 21, 2020. https://en.wikipedia.org/wiki/Bengali_language_movement (accessed Nov. 12, 2024).

[3] B. Baharudin, L. H. Lee, and K. Khan, “A Review of Machine Learning Algorithms for Text-Documents Classification,” Journal of Advances in Information Technology, vol. 1, no. 1, Feb. 2010, doi: https://doi.org/10.4304/jait.1.1.4-20.

[4] I. Ahmad, F. AlQurashi, and R. Mehmood, “Potrika: Raw and Balanced Newspaper Datasets in the Bangla Language with Eight Topics and Five Attributes,” arXiv.org, 2022. https://doi.org/10.48550/arXiv.2210.09389 (accessed Nov. 12, 2024).

[5] Md. Habibullah, Md. Shymon Islam, Fatima Tuz Jahura, and J. Biswas, “Bangla Document Classification Based on Machine Learning and Explainable NLP,” Dec. 2023, doi: https://doi.org/10.1109/eict61409.2023.10427766.

[6] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural Language processing: State of the art, Current Trends and Challenges,” Multimedia Tools and Applications, vol. 82, no. 3, pp. 3713–3744, Jul. 2022, doi: https://doi.org/10.1007/s11042-022-13428-4.

[7] A. D. Chamorro, J. Seguel, and K. S. Ramos, “Transformer-based modeling to study repetitive sequences of the human genome,” *Elsevier eBooks*, pp. 75–82, Sep. 2023, doi: https://doi.org/10.1016/b978-0-12-824010-6.00059-9.

[8] I. Ahmad, F. AlQurashi, and R. Mehmood, “Potrika: Raw and Balanced Newspaper Datasets in the Bangla Language with Eight Topics and Five Attributes,” *arXiv (Cornell University)*, Jan. 2022, doi: https://doi.org/10.48550/arxiv.2210.09389.

[9] F. Alam et al., “A Review of Bangla Natural Language Processing Tasks and the Utility of Transformer Models,” arXiv.org, 2021. https://doi.org/10.48550/arXiv.2107.03844 (accessed Nov. 12, 2024).

[10] A. Hossain, N. Chaudhary, Z. Hasan Rifad, and B. M. M. Hossain, “Bangla News Headline Categorization,” International Journal of Education and Management Engineering, vol. 11, no. 6, pp. 39–48, Dec. 2021, doi: https://doi.org/10.5815/ijeme.2021.06.05.

[11] M. Hasan, L. Islam, I. Jahan, Sabrina Mannan Meem, and R. M. Rahman, “Natural Language Processing and Sentiment Analysis on Bangla Social Media Comments on Russia-Ukraine War using Transformers,” Mar. 2023, doi: https://doi.org/10.1142/s2196888823500021.

[12] T. A. Mahmud, S. Sultana, and A. Mondal, “A New Technique to Classification of Bengali News Grounded on ML and DL Models,” International Journal of Computer Applications, vol. 185, no. 18, pp. 15–21, Jun. 2023, doi: https://doi.org/10.5120/ijca2023922897.

[13] Md. M. Rahman, Md. A. Z. Khan, and A. A. Biswas, “Bangla News Classification using Graph Convolutional Networks,” IEEE Xplore, Jan. 01, 2021. https://ieeexplore.ieee.org/abstract/document/9402567 (accessed Nov. 12, 2024).

[14] J. Karhunen, T. Raiko, and K. Cho, “Unsupervised deep learning,” Advances in Independent Component Analysis and Learning Machines, pp. 125–142, 2015, doi: https://doi.org/10.1016/b978-0-12-802806-3.00007-5.

[15] S. Piduguralla, “Transformers: Revolutionizing Natural Language Processing,” Medium, Jun. 17, 2024. https://medium.com/@tejaswaroop2310/transformers-revolutionizing-natural-language-processing-6509bb109f06 (accessed Nov. 09, 2024).

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv.org, May 24, 2019. https://arxiv.org/abs/1810.04805# (accessed Nov. 12, 2024).

[17] A. Bhattacharjee *et al.*, “BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla,” *Findings of the Association for Computational Linguistics: NAACL 2022*, Jan. 2022, doi: https://doi.org/10.18653/v1/2022.findings-naacl.98.

[18] M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar, and T. Koshiba, “Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding,” IEEE Access, vol. 10, pp. 91855–91870, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3197662.

[19] L. Alzubaidi et al., “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” Journal of Big Data, vol. 8, no. 1, Mar. 2021, doi: https://doi.org/10.1186/s40537-021-00444-8.

[20] J. C. Olamendy, “Backpropagation in Deep Learning: The Key to Optimizing Neural Networks,” Medium, Jul. 15, 2024. https://medium.com/@juanc.olamendy/backpropagation-in-deep-learning-the-key-to-optimizing-neural-networks-7c063a03f677 (accessed Nov. 12, 2024).

[21] S. Edem, “Survey on Recurrent Neural Network in Natural Language Processing,” International Journal of Engineering Trends and Technology - IJETT, 2017. https://ijettjournal.org/archive/ijett-v48p253 (accessed Nov. 12, 2024).

[22] R. Dey and F. M. Salem, “Gate-variants of Gated Recurrent Unit (GRU) neural networks,” 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Aug. 2017, doi: https://doi.org/10.1109/mwscas.2017.8053243.

[23] G. Van Houdt, C. Mosquera, and G. Nápoles, “A review on the long short-term memory model,” Artificial Intelligence Review, vol. 53, no. 8, May 2020, doi: https://doi.org/10.1007/s10462-020-09838-1.

[24] Md. R. Hossain, M. M. Hoque, N. Siddique, and I. H. Sarker, “Bengali text document categorization based on very deep convolution neural network,” Expert Systems with Applications, vol. 184, p. 115394, Dec. 2021, doi: https://doi.org/10.1016/j.eswa.2021.115394.

Thesis_rep

ORIGINALITY REPORT

11%	9%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	Submitted to University of Carthage Student Paper	1%
3	www.geeksforgeeks.org Internet Source	1%
4	Deb, Dipok. "Application and Analysis of Machine Learning and Deep Learning Algorithms in Detection of DDoS Cyberattacks", The University of Texas Rio Grande Valley, 2024 Publication	1%
5	www.dhakatribune.com Internet Source	1%
6	export.arxiv.org Internet Source	<1%
7	Submitted to Daffodil International University Student Paper	<1%
8	link.springer.com Internet Source	