A Deep Learning Approach on Bengali News Classification using Gated Recurrent Neural Network

Faisal Islam[1], Mamun Miah[2]
National University.

ABSTRACT

In this study, we present a Bangla news classification model based on the Gated Recurrent Unit (GRU) architecture. The goal of the model is to automatically categorize Bangla news headlines into predefined classes. News classification is a vital task in information retrieval and recommendation systems, aiding users in accessing relevant news content quickly. The proposed model leverages the sequential nature of natural language by utilizing GRU units to effectively capture temporal dependencies within the news headlines. We preprocess the text data using tokenization and padding to prepare it for training. The model is trained on a labeled dataset containing news headlines and their corresponding categories. We evaluate the model's performance using standard evaluation metrics. The results demonstrate the effectiveness of the GRU-based model in accurately classifying Bangla news headlines into their respective categories, contributing to more efficient and personalized news delivery systems for Bangla-speaking users. The model's ability to handle the contextual nuances of the Bangla language makes it a valuable tool in various real-world applications, such as content recommendation and personalized news aggregation.

ACKNOLEDGEMENT

At first We would like to thank the almighty Allah who created us with not only the ability to design and program this system but also the power of patience.

Our appreciation heartfelt gratitude goes to our project supervisor **Amita Chakroborty,** Associate Professor, Department of Computer Science and Engineering, Shaikh Burhanuddin Post Graduate College. We are obliged and thankful to her for her continuous encouragement, motivation and professional guidance during the work of this project which has proven to be an integral part of it. Without her valuable support and guidance, this project could not elevate up this level of development from our point of view.

Last of all we are grateful to all our teachers who directly or indirectly helped us to complete this project report.

Table of Contents

CHAPTER 06: Implementation and Experiment

CHAPTER 07: Conclusion

1.Introduction

1.1 About Project

The "Bangla News Classification Using GRU Model" project is a focused endeavor aimed at addressing a pressing gap in the realm of Natural Language Processing (NLP) and machine learning. With an extensive population of approximately 284.3 million Bangla speakers worldwide, the project seeks to harness the power of deep learning techniques to efficiently categorize and classify Bangla news headlines into various predefined categories.

At the core of the project lies the utilization of the Gated Recurrent Unit (GRU), a specialized type of Recurrent Neural Network (RNN) designed to capture sequential patterns in data. The GRU's ability to understand the contextual nuances of language makes it an ideal candidate for processing Bangla text, which is rich in grammatical and linguistic complexities.

The project's key objectives encompass several facets. First, it involves the curation and preprocessing of a dataset comprising Bangla news headlines across diverse categories such as "International," "Sports," "Politics," and more. This dataset forms the foundation for training and validating the GRU model.

The GRU model architecture is meticulously crafted to account for the intricacies of Bangla language structure. The input news headlines are tokenized and embedded, preserving their semantic essence. The model is then trained to learn and recognize subtle patterns, linguistic cues, and contextual information that contribute to accurate categorization.

To gauge the model's performance, a comprehensive evaluation is conducted, involving metrics such as accuracy, precision, recall, and F1-score. The validation process ensures that the model generalizes well to unseen data, upholding its reliability and predictive capabilities.

The potential impact of this project is significant. Accurate and automated classification of Bangla news headlines holds the promise of revolutionizing the news consumption experience for millions of Bangla speakers. It aids in disseminating information efficiently, enhances user engagement by providing tailored news recommendations, and empowers individuals to access credible news sources in their preferred language.

1.2 Machine Learning (ML)

Machine Learning (ML) is a transformative field within the realm of artificial intelligence that empowers computers to learn from data and improve their performance over time. Rooted in the idea that machines can learn patterns and make decisions autonomously, ML has rapidly evolved to become an integral part of various industries and applications, fundamentally reshaping how we interact with technology and harness the power of data.

At its core, ML revolves around the concept of algorithms and models that learn from examples. Rather than being explicitly programmed to perform specific tasks, ML systems are designed to autonomously improve their performance by learning from data inputs. This data-driven approach allows machines to uncover complex patterns, generate insights, and make informed decisions based on the information they have been exposed to.

1.3 Neural Network

A neural network is a computational model inspired by the structure and functioning of the human brain. It is a fundamental concept in the field of artificial intelligence and machine learning. Neural networks consist of interconnected nodes, also known as "neurons," that work collectively to process and analyze data, recognize patterns, and make predictions. These networks are capable of learning from examples, adjusting their parameters based on input data, and improving their performance over time.

Key Components of a Neural Network:

Neurons (Nodes): Neurons are the fundamental processing units in a neural network. Each neuron receives input signals, performs a computation, and produces an output signal. Neurons are organized in layers, with each layer having a specific role in data transformation.

Weights and Bias: Neurons are connected by weighted connections, which determine the strength of the influence one neuron has on another. Each connection has an associated weight that adjusts during training to optimize the network's performance. A bias term is also often used to shift the activation function of a neuron.

Activation Function: An activation function defines the output of a neuron based on its input. It introduces non-linearity into the network, allowing it to capture complex relationships in the data. Common activation functions include sigmoid, ReLU (Rectified Linear Unit), and tanh (hyperbolic tangent).

Layers: A neural network consists of one or more layers of neurons. The input layer receives raw data, the hidden layers process and transform the data, and the output layer produces the final prediction or classification.

1.4 History of Neural Network

The history of neural networks has been marked by distinct phases of progress, stagnation, and resurgence. Its origins trace back to the 1940s and 1950s when Warren McCulloch and Walter Pitts introduced the concept of artificial neurons, laying the groundwork for the neural network model. The 1960s and 1970s saw the development of the perceptron by Frank Rosenblatt, a significant advancement in its time, yet its limitations became evident as it struggled with complex tasks.

Subsequently, the field faced challenges during the "AI winter" of the 1980s and 1990s, where neural network research encountered skepticism and stagnation due to computational limitations and training complexities. However, a breakthrough arrived in 1986 with the introduction of the backpropagation algorithm, enabling more efficient training of multi-layer perceptrons and reigniting interest in neural networks.

In 1989, Yann LeCun's convolutional neural networks (CNNs) marked a significant turning point, revolutionizing image recognition and processing. Despite this progress, neural networks faced competition from alternative algorithms like support vector machines in the 1990s and 2000s.

The landscape shifted again in the mid-2000s with the advent of the deep learning resurgence. Fueled by increased computational power and the availability of vast datasets, researchers began training deep neural networks with multiple layers, achieving remarkable breakthroughs in image and speech recognition. Notable architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks addressed the vanishing gradient problem, propelling the field forward.

In the 2010s, neural networks achieved new heights. Convolutional neural networks demonstrated human-level performance in image classification, while recurrent neural networks made significant strides in natural language processing tasks. This decade also witnessed innovations such as generative adversarial networks (GANs) for image generation and transformers with attention mechanisms for language understanding.

In recent years, neural networks have become a cornerstone of artificial intelligence research, playing a pivotal role in various applications across industries. This historical trajectory showcases the resilience of researchers and the profound impact of technological advancements, solidifying neural networks as a fundamental component of modern machine learning and AI.

1.5 Expected Output

The expected output of a Bangla news headline categorization project would be a model that can accurately classify Bangla news headlines into predefined categories. The model takes a Bangla news headline as input and assigns it to one of the predefined categories based on the content of the headline. Here's how the project's output might look:

Input: A Bangla news headline text.

Output: The model's predicted category for the input news headline. This output will be the category label or name that corresponds to the category the model believes the headline belongs to.

For example:
Input: "বাংলাদেশ জয়ের প্রত্যুষা"
Output: "Sports"
Or,
Input: "প্রধানমন্ত্রী বাংলাদেশের প্রয়াস সরাসরি নিরীক্ষণ করেন"
Output: "Politics"

Our model's success will be evaluated based on how accurately it can predict the correct category for a given news headline. You would typically assess the model's performance using metrics like accuracy, precision, recall, and F1-score on a validation or test dataset that the model hasn't seen during training.

Overall, the goal of the project is to develop a model that can assist in automatically categorizing Bangla news headlines, which can be beneficial for news organizations, websites, and analysis of trends in the Bangla news landscape.

1.6 Aims and Objectives

The aim of this project is to develop a machine learning model for accurately categorizing Bangla news headlines into predefined categories. The project aims to leverage natural language processing (NLP) techniques to automate the process of news categorization, thereby improving the efficiency of news organizations and enhancing the user experience for readers.

Preprocess the Bangla news headlines dataset to prepare it for machine learning. Design and implement a suitable machine learning algorithm for text classification. Train the model using the preprocessed dataset and evaluate its performance using appropriate metrics. Fine-tune the model's hyper parameters to achieve optimal accuracy and efficiency.

2.Literature Review

2.1 Introduction

The Bangla language, spoken by millions across the globe, is an integral part of the cultural fabric of South Asia. With its rich linguistic nuances and distinct grammatical structure, it presents both opportunities and complexities in the realm of Natural Language Processing (NLP). As news headlines span a multitude of topics encompassing politics, sports, international affairs, and more, the task of accurately classifying them demands a sophisticated approach. By integrating the power of neural networks, this project endeavors to build a robust model capable of deciphering the intricate textual characteristics that define news categories. Neural networks, known for their capacity to learn complex patterns and relationships within data, stand poised to unlock the hidden insights within Bangla news headlines. The versatility of neural network architectures, including variants like Gated Recurrent Units (GRUs) and Convolutional Neural Networks (CNNs), empowers the model to comprehend the contextual nuances and semantic subtleties that distinguish different news topics.

2.2 Background Study

Many researchers have done a lot of research on the classification of Bengali News, and neural networks have been used a lot in these studies along with many other methods.

In this part, we will discuss different types of research which are related to the classification of Bengali News.

In a paper [1], the author has done a classification of Bengali news tags and titles. They used CNN method to reach their final output and in order to extract characteristics, there were two parallel branches which gave them 93.47% accuracy. As they were unable to gather any newly updated news, they produced a set of data with 88,968 news items. They desire to use the GAN approach in future projects.

The authors of the research [2] classified different news stories in Bengali. They compared several machine and deep learning techniques for this purpose. They divided the data into manual and automatic labels where they found that with the collaboration of Gated Recurrent Unit (GRU) algorithm and Fasttext approach for embedding words have the highest accuracy of 91.83% for manual labels. On the other hand, K-Nearest Neighbor (KNN) strategy and Doc2Vec word embedder generated two outcomes for automated categorization that are too low. They want to use hybrid techniques for deep learning later on along with a variety of embedded word approaches.

In the study [3], authors have tried to classify bengali news with the help of different kinds of methods where they used 5 different types of machine learning approaches and 2 neural network approaches where they compared among them. But they did not get any good results from their models. Their maximum accuracy was produced by Long Short Term Memory (LSTM) method of neural network which was 87%. For further study, authors want to work with a diverse range of data and try to improve their system.

Rahman et al. [4] proposed an approach which can categorize Bengali news. For this work, authors used data from an online news website named Prothom Alo. In this study they applied many algorithms for different models and compared them. Finally they have achieved a good accuracy using Text-Graph Convolutional Network (Text-GCN) technique which was 96.25%. But still there are some storage problems in their work and they want to solve these difficulties for further process.

A recent study [5] proposed a technique to segment Bengali news headlines into different categories. To build their model they have used Bi-directional Gated Recurrent Unit (BiGRU) approach and compared it with several other models which had already been prepared in various studies. The performance of the given model on data from validation was 84% accurate. They intend to create a Web API for Bengali news in the future.

In the paper [6], researchers have shown various models to divide the Bengali news into multiple segments according to their particular types. For this research, they have used some machine learning and neural network algorithms and to train their model they collected two dataset of news. Their first set of data were gathered from a well-known newspaper Prothom Alo and the second one from Kaggle. At the end of their work, they found that the Multi-layer dense neural network model performed best which achieved 95.50% accuracy. They want to upgrade their model that's why they will apply several approaches to neural networks in their further work.

In the study [7], authors have tried to develop a new model to segment the several news by the type of that news which were in Bengali language. To build a good model they applied various algorithms from deep learning and machine learning approaches. Their dataset was collected from different newspapers and they converted their data into news text. Finally they achieved 93.43% accuracy by CNN model which was higher than other models.

Researchers tried to develop an efficient model to classify Bengali text into 12 segments based on text type in the study [8]. For this reason, they have done their work with few deep learning methods. But they didn't get a good result. As their final outcome, they found that Hybrid CNN–LSTM technique gave the maximum accuracy which was 88.56% for the 10 types and 84.93% for the other all. They have many problems with their dataset which was not sufficient and sound. That's why again they want to do the same types of work with a better dataset and various ways.

In the study [9], authors proposed a technique to divide the Bengali text documents into their category and for this work they presented RNN algorithm with BiLSTM. This paper also compared their presented model with others different type of method like KNN, SVM, Naive Bayes, etc. After completing their work they noticed that the RNN method has been able to gain a 98.33% accuracy rate which was so good. Authors have planned to build a web API and selected TF-IDF based NN in their further task.

In the paper [10], researchers have categorized Bengali articles into 4,302 labels with the help of ML-KNN and Neural Network approaches which were supervised methods. They collected their data from Prothom Alo newspaper. They have measured Precision, Recall and F1-Score for different sectors, but the results which were found by them, were not satisfactory. Here they used Count-Vectorizer for embedding words, but in future they want to apply various methods for this task. For this sort of issue, they believed a technique known as unsupervised learning would be a more practical answer.

3.Navigating Complexity: Understanding Neural Network Complexity

3.1 Neural Network Architecture

A neural network architecture is the blueprint or structure that defines how artificial neural networks are organized. Neural networks are computational models inspired by the human brain's neural networks, designed to learn patterns and relationships from data. The architecture dictates how neurons (nodes) are connected and how information flows through the network.

Here are some key components and concepts related to neural network architecture:

Neurons (Nodes): Neurons are the basic computational units in a neural network. They receive input, perform a weighted sum and activation function operation, and pass the output to the next layer.

Layers: A neural network is composed of layers. The main types of layers are:

i. Input Layer: The initial layer that receives raw data.

ii. Hidden Layers: Intermediate layers between the input and output layers. They learn hierarchical representations of the data.

iii. Output Layer: The final layer that produces the network's predictions.

Connections (Weights): Neurons are connected with weighted connections that represent the strength of the connection. These weights are learned during the training process to optimize the network's performance.

Activation Functions: Activation functions introduce non-linearity to the network, enabling it to learn complex relationships. Common activation functions include ReLU (Rectified Linear Unit), sigmoid, and tanh.

Feedforward Architecture: In a feedforward neural network, data flows from the input layer through the hidden layers to the output layer. There are no loops or cycles in this architecture.

Recurrent Architecture: Recurrent Neural Networks (RNNs) have connections that allow loops to exist in the network. This enables them to process sequences of data, making them suitable for tasks like natural language processing and time series analysis.

Convolutional Architecture: Convolutional Neural Networks (CNNs) are designed for processing grid-like data, such as images. They use convolutional layers to automatically learn features from input data.

Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU): These are specialized types of recurrent layers that address the vanishing gradient problem and enable RNNs to learn longer-term dependencies in sequences.

Architecture Design: Designing a neural network architecture involves selecting the number of layers, the number of neurons in each layer, activation functions, regularization techniques, and more. This design process can be guided by domain knowledge, experimentation, and trial and error.

Hyper parameters: Architecture-related settings that are not learned during training are called hyper parameters. These include learning rate, batch size, number of layers, and more.

Deep Learning: Neural networks with multiple hidden layers are known as deep neural networks. Deep learning leverages these deep architectures to learn complex representations from data.

Transfer Learning: Pre-trained neural network architectures, such as VGG, ResNet, and BERT, have already learned meaningful representations from vast amounts of data. Transfer learning involves using these architectures as a starting point for specific tasks.

3.2 Neural Network Architecture Types of neural network

Neural networks come in various types, each designed to tackle specific types of problems and data structures. Here's an overview of some common types of neural network architectures:

i. Feedforward Neural Network (FNN):

- Also known as a Multilayer Perceptron (MLP).
- Consists of an input layer, one or more hidden layers, and an output layer.
- Each neuron in a layer is connected to every neuron in the subsequent layer.
- Used for various tasks including classification, regression, and pattern recognition.

ii. Convolutional Neural Network (CNN):

- Primarily designed for processing grid-like data such as images.
- Utilizes convolutional layers to automatically learn features from input data.
- Employs pooling layers to downsample and reduce spatial dimensions.
- Effective for image classification, object detection, and image generation.

iii. Recurrent Neural Network (RNN):

- Designed for sequence data where the order of elements matters (e.g., time series, natural language).
- Contains loops to allow information to be passed from one step of the network to the next.
- Suffers from vanishing gradient problem in long sequences.
- Extensions like LSTM and GRU were developed to address the vanishing gradient problem.

iv. Long Short-Term Memory (LSTM):

- A specialized type of RNN that can handle long-range dependencies and vanishing gradient issues.
- Contains memory cells that can store information over long periods.
- Widely used in tasks involving sequences, such as language modeling and machine translation.

v. Gated Recurrent Unit (GRU):

- Another type of RNN designed to address vanishing gradient and long-range dependency issues.
- Similar to LSTM but with fewer parameters, making it computationally more efficient.
- Used in tasks like language modeling, speech recognition, and video analysis.

vi. Auto encoder:

- Consists of an encoder that compresses data into a latent representation and a decoder that reconstructs the original data.
- Used for dimensionality reduction, feature learning, and generative tasks.
- Variational Autoencoders (VAEs) introduce probabilistic encoding for generative modeling.

vii. Generative Adversarial Network (GAN):

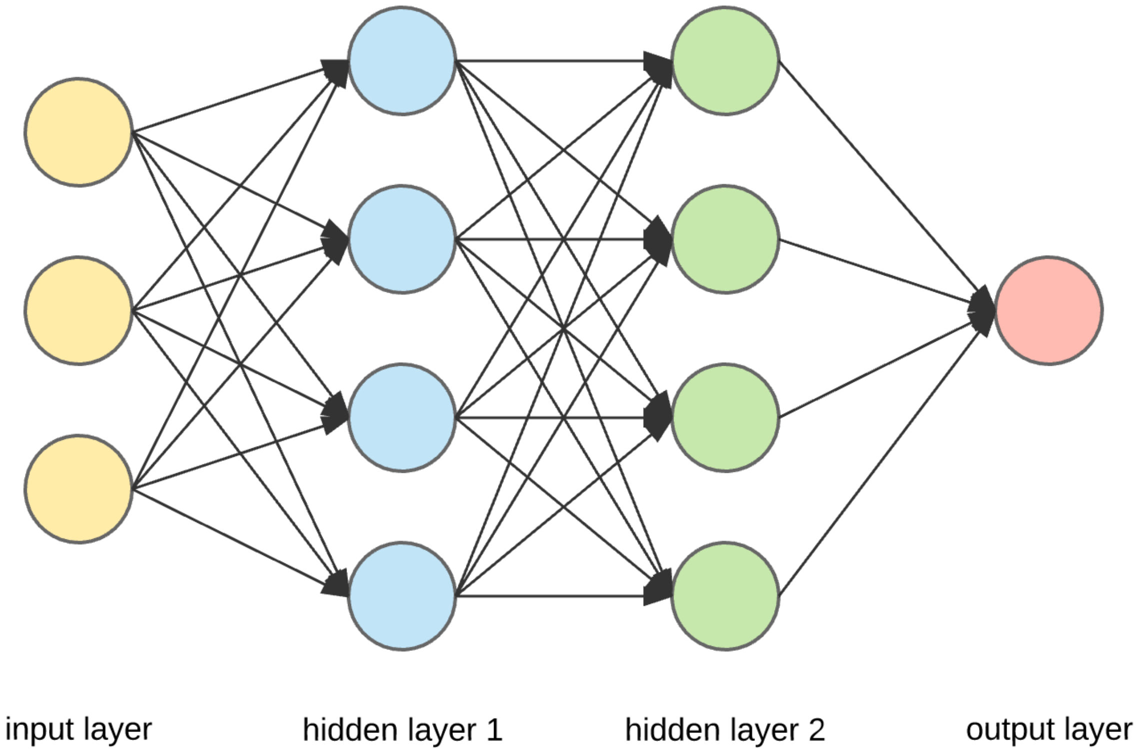
- Comprises a generator network and a discriminator network in a competitive setup.
- The generator aims to create realistic data, while the discriminator aims to differentiate real from generated data.
- Used for generating realistic images, style transfer, and data augmentation.

viii. Transformer:

- Introduced in the context of natural language processing.
- Uses self-attention mechanisms to capture relationships between words in a sequence.
- Popularized by models like BERT and GPT, achieving state-of-the-art results in various NLP tasks.

ix. Siamese Network:

- Designed to learn similarity between inputs by using two identical subnetworks.
- Often used in tasks like face verification, signature verification, and similarity-based recommendation.



3.3 Complexity of Neural Network

The complexity of a neural network refers to the computational and memory resources required for training and inference. It can be influenced by factors such as the number of layers, neurons, parameters, and the data used. Here are some aspects that contribute to the complexity of neural networks:

Model Size:

- The number of parameters in the network affects its size and memory requirements.
- Larger networks with more parameters can capture intricate patterns but require more memory for storage.

Depth and Width:

- Deeper networks (more layers) can capture more complex features but may be prone to vanishing gradients during training.
- Wider networks (more neurons per layer) can learn fine-grained details but demand more computation.

Input and Output Size:

- Networks dealing with high-dimensional inputs, such as images or sequences, tend to be more complex due to the increased number of connections.

Batch Size:

- Larger batch sizes during training can lead to faster convergence but require more memory.
- Smaller batch sizes may lead to slower convergence and less efficient hardware utilization.

Activation Functions:

- Non-linear activation functions introduce complexity, enabling networks to capture complex relationships in data.

Regularization Techniques:

- Techniques like dropout and weight regularization add complexity but can prevent overfitting.

Special Layers:

- Layers like attention mechanisms (as in Transformers) or memory cells (as in LSTMs) introduce additional complexity.

Data Augmentation:

- Complex data augmentation can increase the diversity of the training data, improving generalization but also requiring more computation.

Preprocessing:

- Complex preprocessing, such as advanced tokenization or embedding techniques, can impact both training time and memory requirements.

Model Architecture:

- Unique architectures, such as GANs, autoencoders, and complex attention mechanisms, introduce specific types of complexity.

Hardware and Framework:

- Different hardware (CPUs, GPUs, TPUs) and deep learning frameworks (TensorFlow, PyTorch) have varying levels of efficiency and support for optimizing neural network computations.

Inference Latency:

- For real-time applications, the inference complexity is crucial. Complex networks might be slower to make predictions.

4.Challenges and Model Analysis

4.1 Data collection

Data collection was one of the hardest part of this project. First we had a plan to collect data from various news website using web scrapping technique. Then its came into picture that most of the site do not permit to scrape their site. Many of them considered it as illegal. So after facing this we move to kaggle to find the ready-made dataset. So we collected our dataset from kaggle. It was also very challenging as there is very few data about bengali newspaper headlines. Some of them have very fewer category and also fewer data. So finally we were able to find a dataset according to our need.

4.2 Model selection

Selecting the right model is a pivotal step in developing an effective Bangla news headline categorization system. This prompt delves into the process of model selection, outlining the considerations and rationale behind choosing the Gated Recurrent Unit (GRU) as the preferred model for this task.

1. Exploring Model Options: Begin by exploring a spectrum of neural network architectures suitable for text classification tasks. Investigate options such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). Evaluate each model's suitability in terms of handling sequential data and capturing nuanced language patterns.

2. Rationale for Choosing GRU: While all considered models possess strengths, the GRU stands out as a robust choice for Bangla news headline categorization for the following reasons:

- Long-Range Dependencies: GRU, an advanced type of RNN, excels at capturing long-range dependencies in sequential data. As news headlines often entail varying sentence lengths and intricate context, the GRU's ability to retain relevant information over extended sequences becomes invaluable.
- Efficiency: GRU strikes a balance between LSTM's complexity and standard RNN's limitations. It contains fewer parameters than LSTM, making it computationally efficient while maintaining the capability to capture important features.
- Vanishing Gradient Mitigation: The GRU's architecture incorporates mechanisms that mitigate the vanishing gradient problem, ensuring effective learning even in deep networks and prolonged sequences.
- Sequential Insights: News headlines convey information progressively. The GRU's recurrent nature facilitates the capture of sequential insights, enabling the model to discern context nuances and refine predictions.
- Effective Text Understanding: GRU's recurrent connections enable it to develop a richer understanding of text, considering the evolving context word by word. This quality is particularly advantageous for analyzing news headlines with varying linguistic structures.

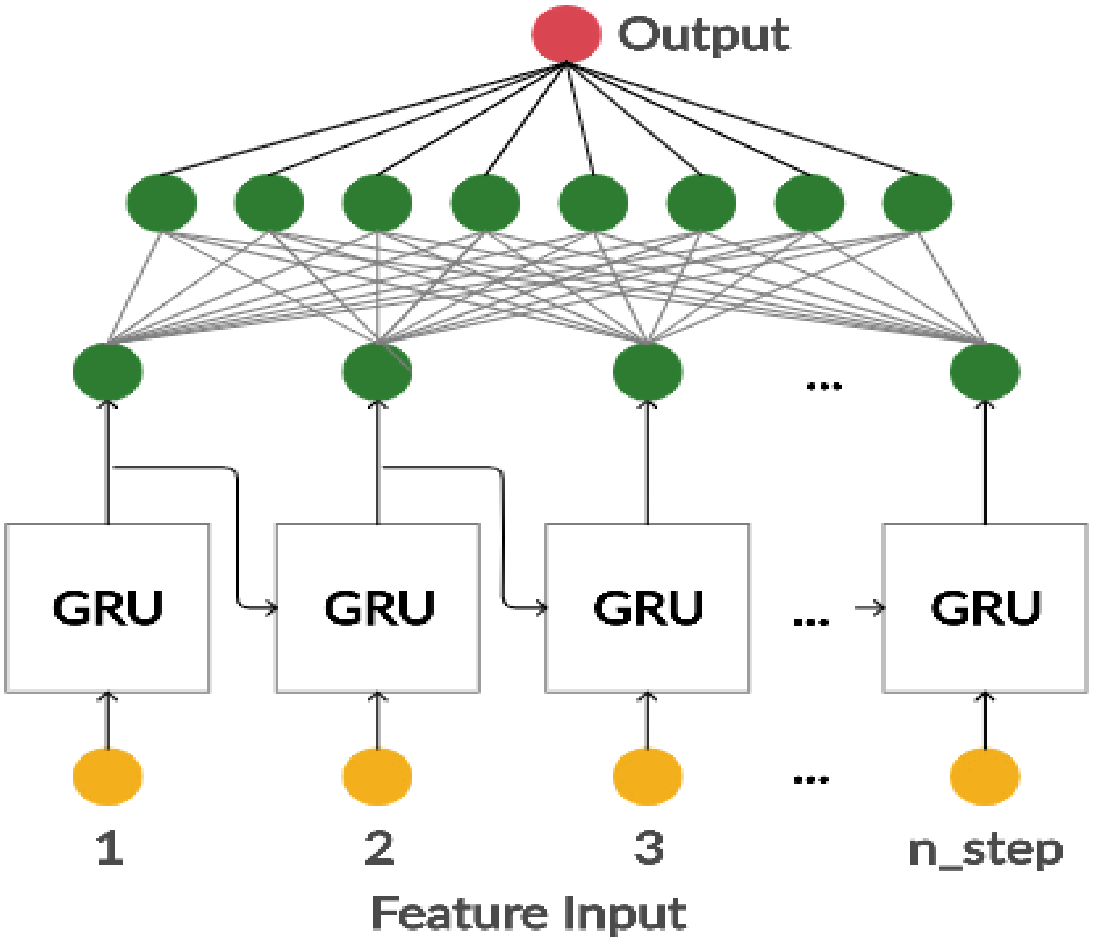
3. Experimental Validation: Conduct empirical experiments to validate the suitability of the GRU model. Train the GRU-based model on a representative dataset of Bangla news headlines, utilizing appropriate preprocessing techniques. Compare the model's performance against alternative architectures using relevant evaluation metrics, such as accuracy, precision, recall, and F1-score.

4. Hyper parameter Tuning: Fine-tune the GRU model's hyper parameters to optimize its performance further. Experiment with parameters like learning rate, batch size, and dropout rates to strike the ideal balance between under fitting and overfitting.

5. Robustness and Scalability: Assess the GRU model's robustness by subjecting it to varied headline types, including both straightforward and nuanced examples. Additionally, evaluate the model's scalability to handle increasing data volumes as well as real-time categorization demands.

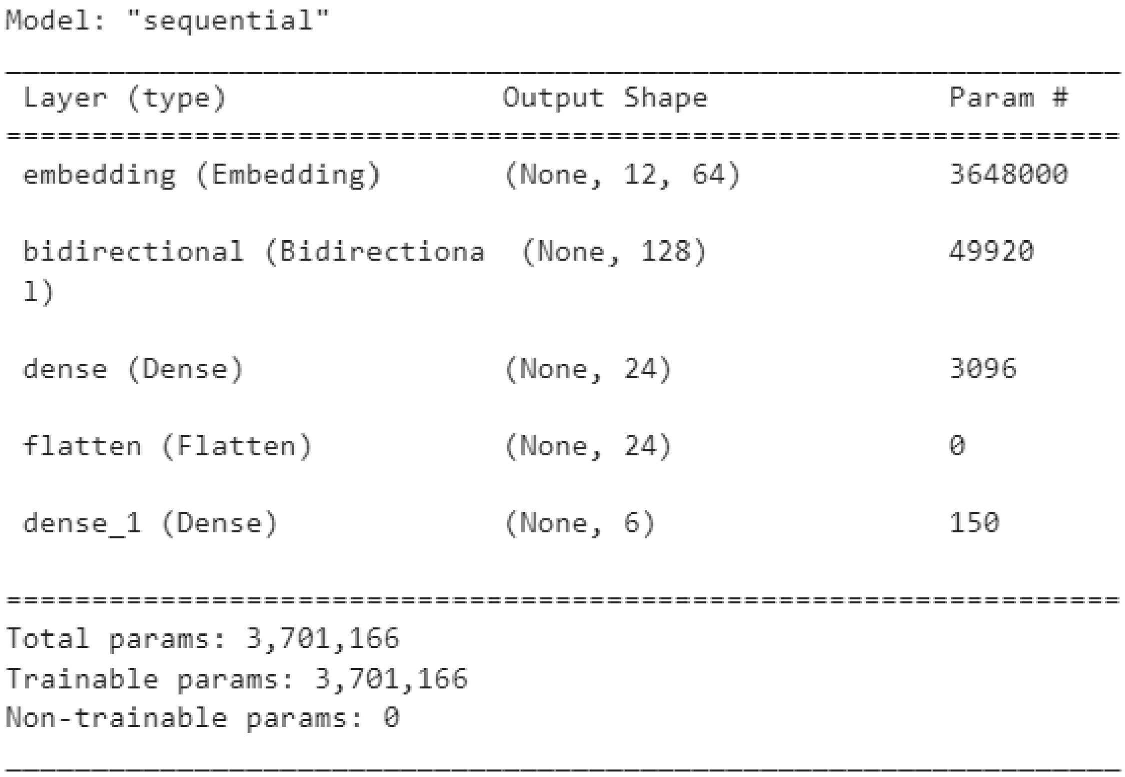
The GRU model emerges as a compelling choice for Bangla news headline categorization due to its adeptness at capturing sequential dependencies, computational efficiency, and the ability to mitigate vanishing gradient challenges. By selecting the GRU model and conducting rigorous experimentation, we aim to harness its potential to accurately categorize Bangla news headlines, contributing to enhanced news organization workflows and insights.

4.3 Model Architecture



4.4 Model Development

Model Development The used model architecture consists of a embedding layer (input_length = 12, embedding_dim = 64), GRU layer (n_units = 64), two dense layer (n_units = 24, 6), a dropout and a softmax layer. The Architecture looks like-



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 12, 64)            3648000

 bidirectional (Bidirectiona  (None, 128)              49920
 l)

 dense (Dense)               (None, 24)                3096

 flatten (Flatten)           (None, 24)                0

 dense_1 (Dense)             (None, 6)                 150

=================================================================
Total params: 3,701,166
Trainable params: 3,701,166
Non-trainable params: 0
_________________________________________________________________
```

5.Research Methodology

5.1 Description

The journey begins with "Data Acquisition," involving the collection of pertinent Bangla news headlines from diverse sources. These headlines serve as the foundation for analysis and insights. Subsequently, the acquired data undergoes "Data Preprocessing," a pivotal phase that ensures data quality and prepares it for subsequent stages of the project. The progression leads to "Data Tokenization," a critical process that converts the news headlines into a structured format. This transformation is fundamental as it allows the neural network model to comprehend and process the textual content effectively. The crux of the project lies in "Training the Model with Training Dataset." In this phase, the neural network is meticulously configured and trained using the tokenized headlines. The architecture of the neural network is carefully designed to decipher intricate patterns and relationships encoded within the news data. Following successful training, the model advances to "Evaluation," a critical checkpoint where its accuracy is rigorously assessed. A dedicated validation dataset is employed to test the model's proficiency in accurately categorizing news headlines into their respective categories. The culmination of the project is embodied in the "Output" stage. Here, the trained neural network stands equipped to receive new and unseen Bangla news headlines, and with remarkable accuracy, predict the precise category they belong to. This ultimate achievement demonstrates the tangible impact of neural network technology in enhancing the news consumption experience for the Bangla-speaking community.

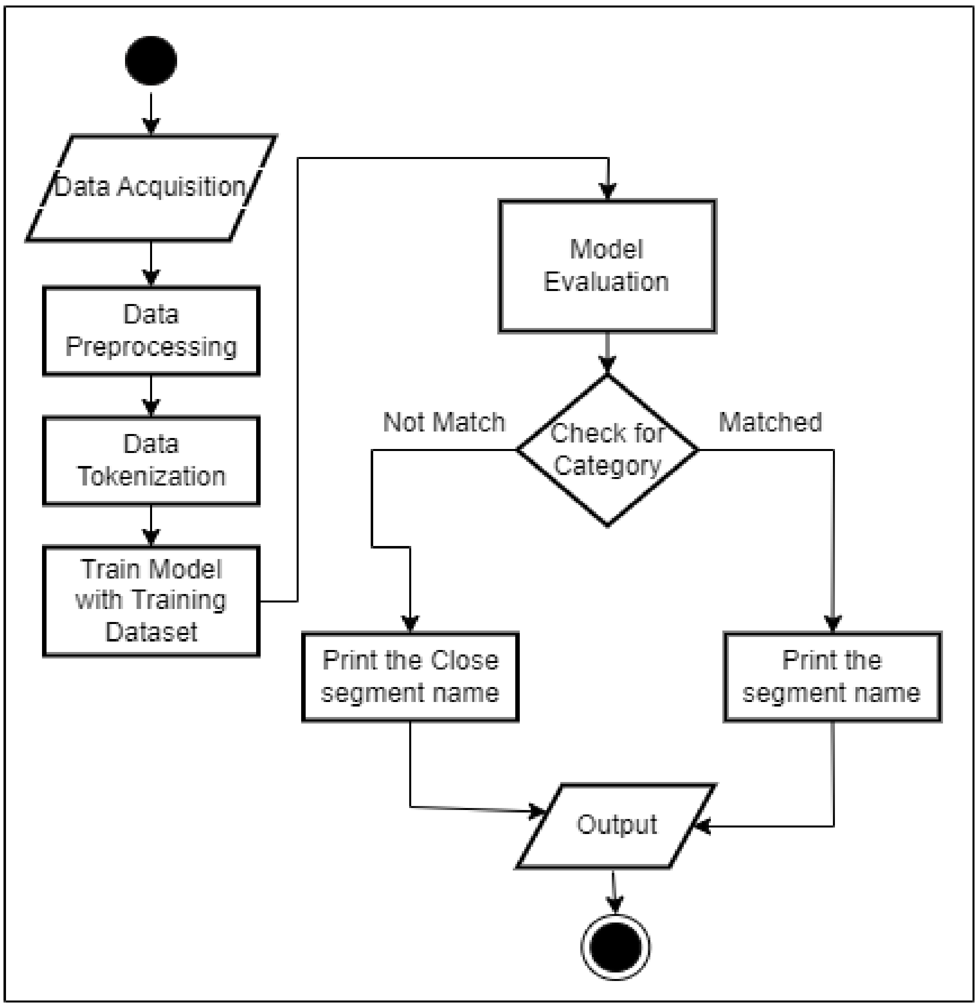
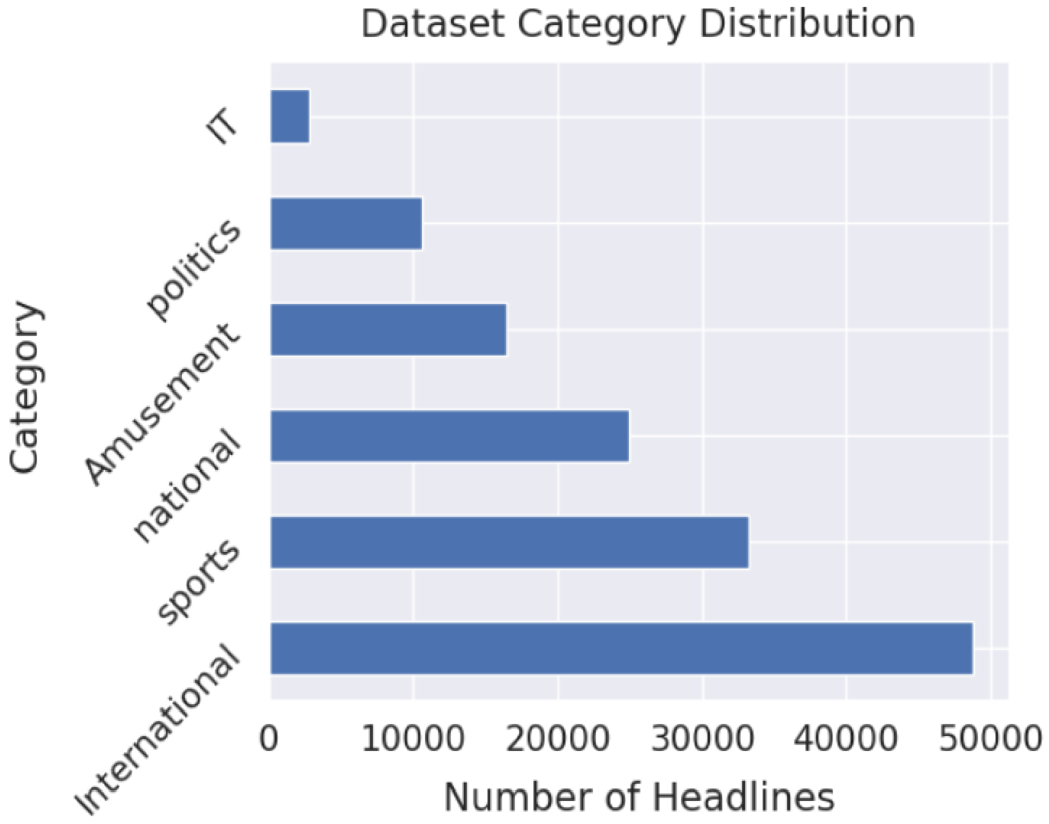


Figure: Flowchart

5.2 Data acquisition

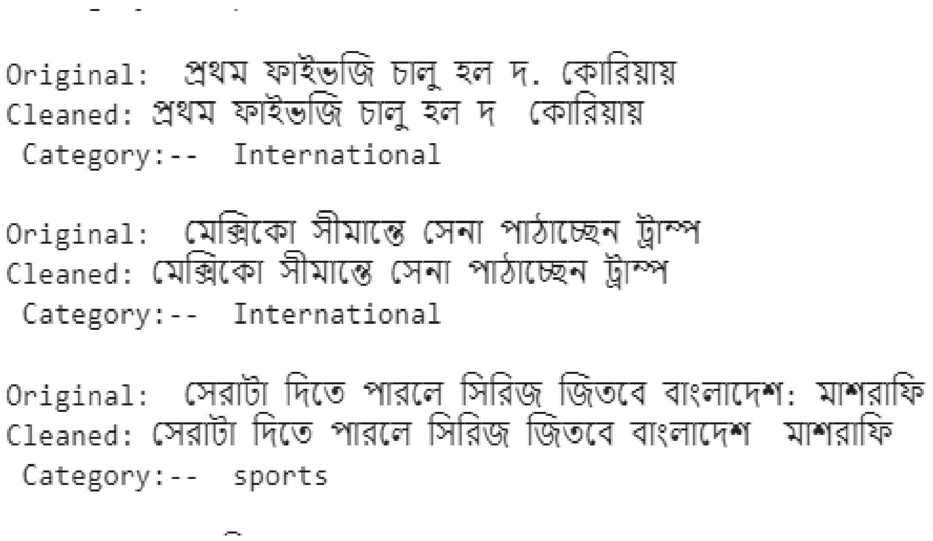
Here we have collected data that contain of 6 classes and total 1,36,811 of data. This category are consist of International, National, Sports, Amusement, Politics and IT.

```
Total number of headlines: 136811
```



5.3 Data Preprocessing

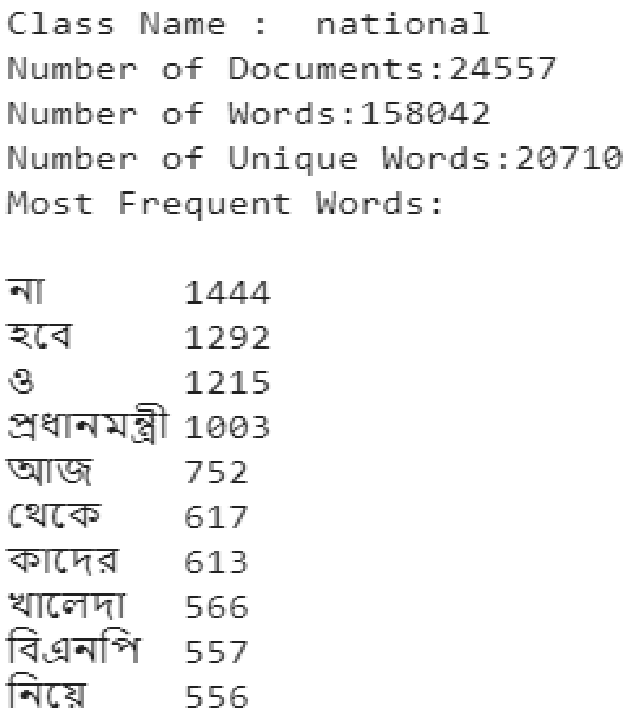
We preprocessed the dataset by dropping the minimum length of headline and also used the unicode of Bangla language to remove the unnecessary marks and sign using the regex expression.



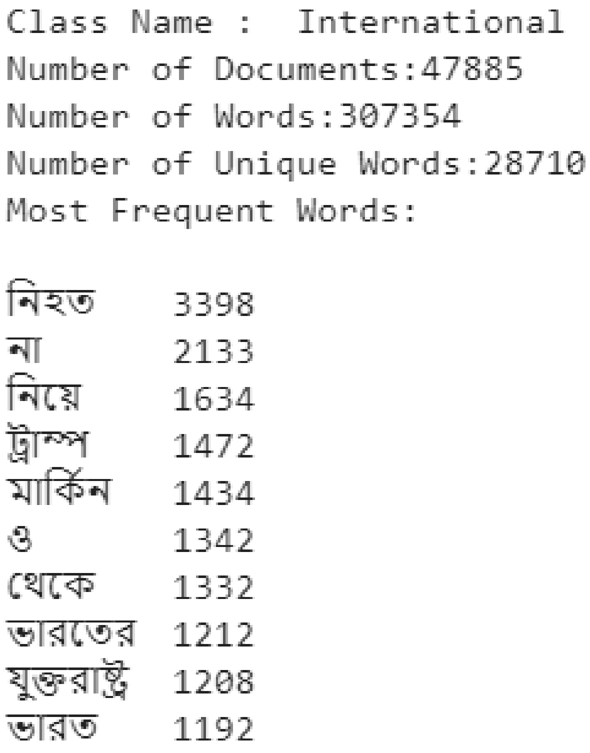
```
Original:  প্রথম ফাইভজি চালু হল দ. কোরিয়ায়
Cleaned: প্রথম ফাইভজি চালু হল দ  কোরিয়ায়
 Category:--  International

Original:  মেক্সিকো সীমান্তে সেনা পাঠাচ্ছেন ট্রাম্প
Cleaned: মেক্সিকো সীমান্তে সেনা পাঠাচ্ছেন ট্রাম্প
 Category:--  International

Original:  সেরাটা দিতে পারলে সিরিজ জিতবে বাংলাদেশ: মাশরাফি
Cleaned: সেরাটা দিতে পারলে সিরিজ জিতবে বাংলাদেশ  মাশরাফি
 Category:--  sports
```

5.3.1 Data Summery

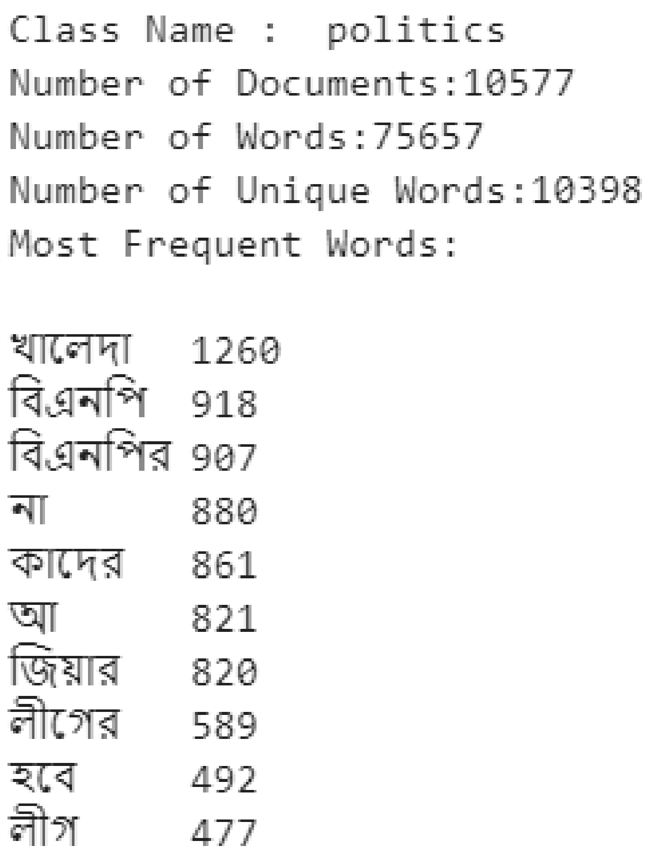
Data summary includes the information about number of documents, words and unique words have in each category class. Also, include the length distribution of the headlines in the dataset.



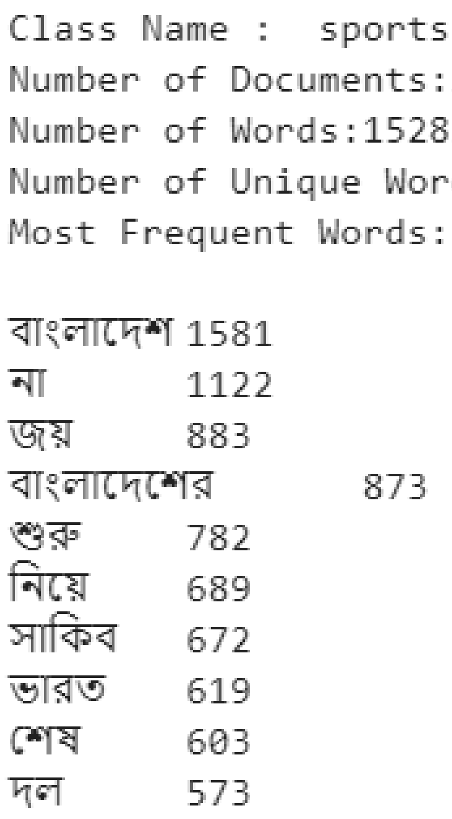
```
Class Name :  national
Number of Documents:24557
Number of Words:158042
Number of Unique Words:20710
Most Frequent Words:

না       1444
হবে      1292
ও        1215
প্রধানমন্ত্রী 1003
আজ       752
থেকে     617
কাদের    613
খালেদা   566
বিএনপি   557
নিয়ে     556
```



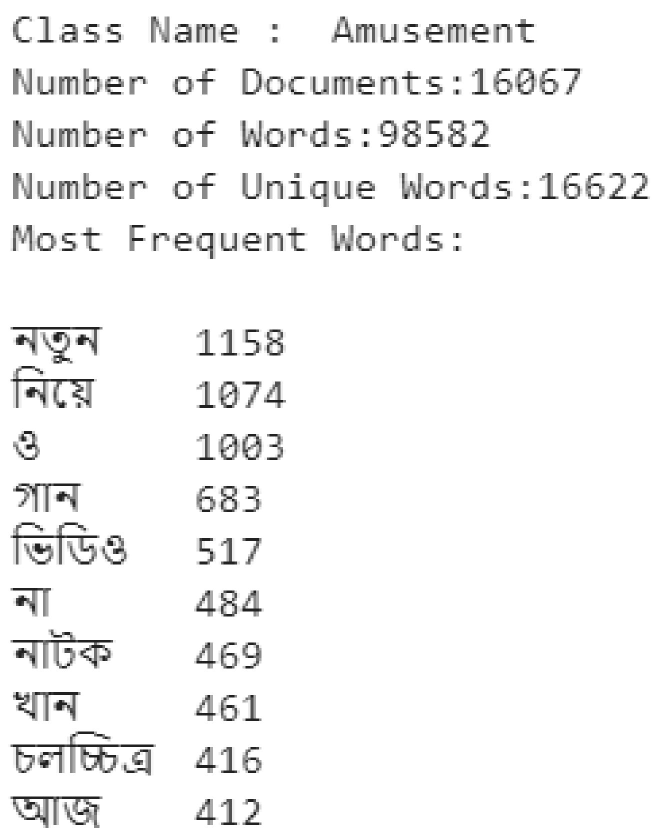
```
Class Name :  International
Number of Documents:47885
Number of Words:307354
Number of Unique Words:28710
Most Frequent Words:

নিহত     3398
না       2133
নিয়ে     1634
ট্রাম্প   1472
মার্কিন   1434
ও        1342
থেকে     1332
ভারতের   1212
যুক্তরাষ্ট্র 1208
ভারত     1192
```



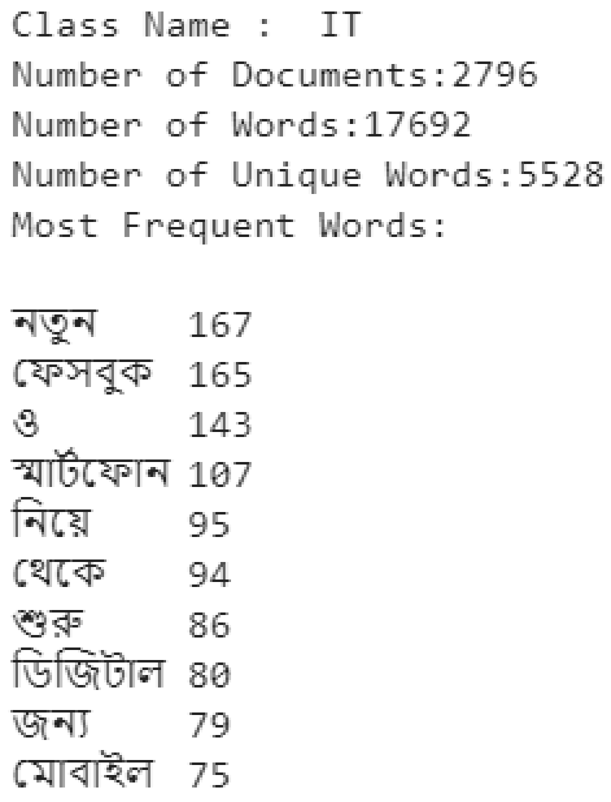
```
Class Name :  politics
Number of Documents:10577
Number of Words:75657
Number of Unique Words:10398
Most Frequent Words:

খালেদা   1260
বিএনপি  918
বিএনপির 907
না       880
কাদের   861
আ       821
জিয়ার   820
লীগের   589
হবে      492
লীগ      477
```



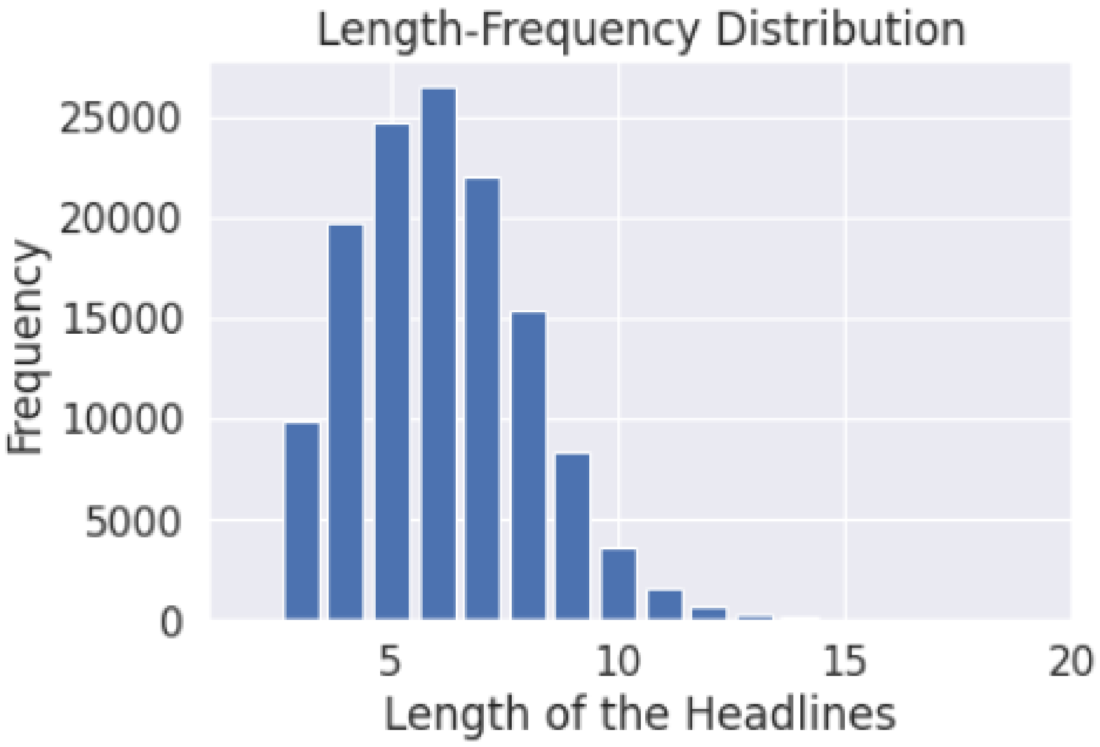
```
Class Name :  sports
Number of Documents:30831
Number of Words:152852
Number of Unique Words:18581
Most Frequent Words:

বাংলাদেশ 1581
না       1122
জয়      883
বাংলাদেশের        873
শুরু     782
নিয়ে     689
সাকিব   672
ভারত    619
শেষ     603
দল      573
```

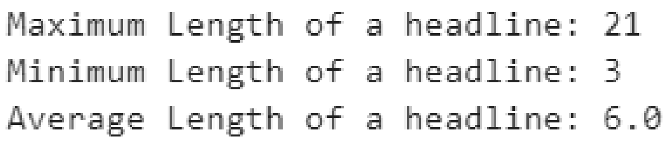


```
Class Name :  Amusement
Number of Documents:16067
Number of Words:98582
Number of Unique Words:16622
Most Frequent Words:

নতুন      1158
নিয়ে      1074
ও         1003
গান       683
ভিডিও     517
না        484
নাটক      469
খান       461
চলচ্চিত্র  416
আজ        412
```



```
Class Name :  IT
Number of Documents:2796
Number of Words:17692
Number of Unique Words:5528
Most Frequent Words:

নতুন      167
ফেসবুক    165
ও         143
স্মার্টফোন 107
নিয়ে      95
থেকে      94
শুরু      86
ডিজিটাল   80
জন্য      79
মোবাইল    75
```





```
Maximum Length of a headline: 21
Minimum Length of a headline: 3
Average Length of a headline: 6.0
```

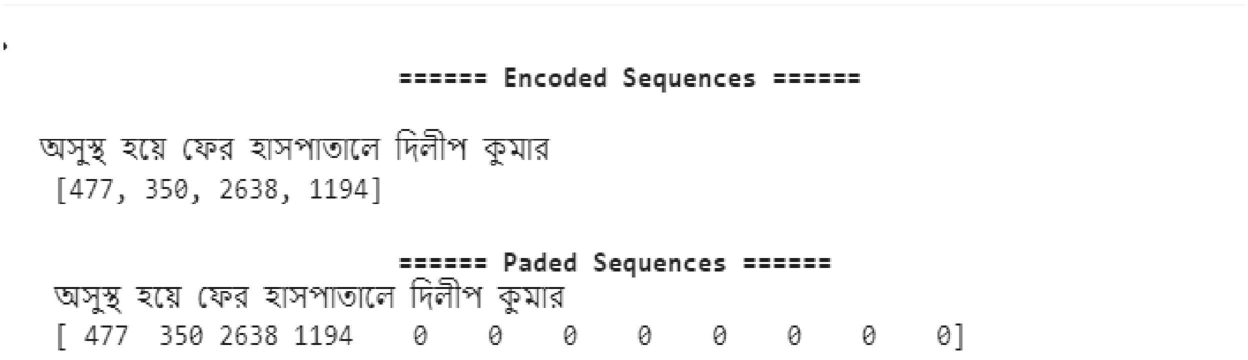
From this graphical information, we can select the suitable length of headlines that we have to use for making every headlines into a same length.

5.4 Data Tokenization

Data tokenization are the most crucial part of this project. As it is text based data so we have to tokenize the dataset so that the each word are assigned with a certain type of numerical encoding. Text to sequence and padded sequence are done respectively as a prt of preprocessing data.

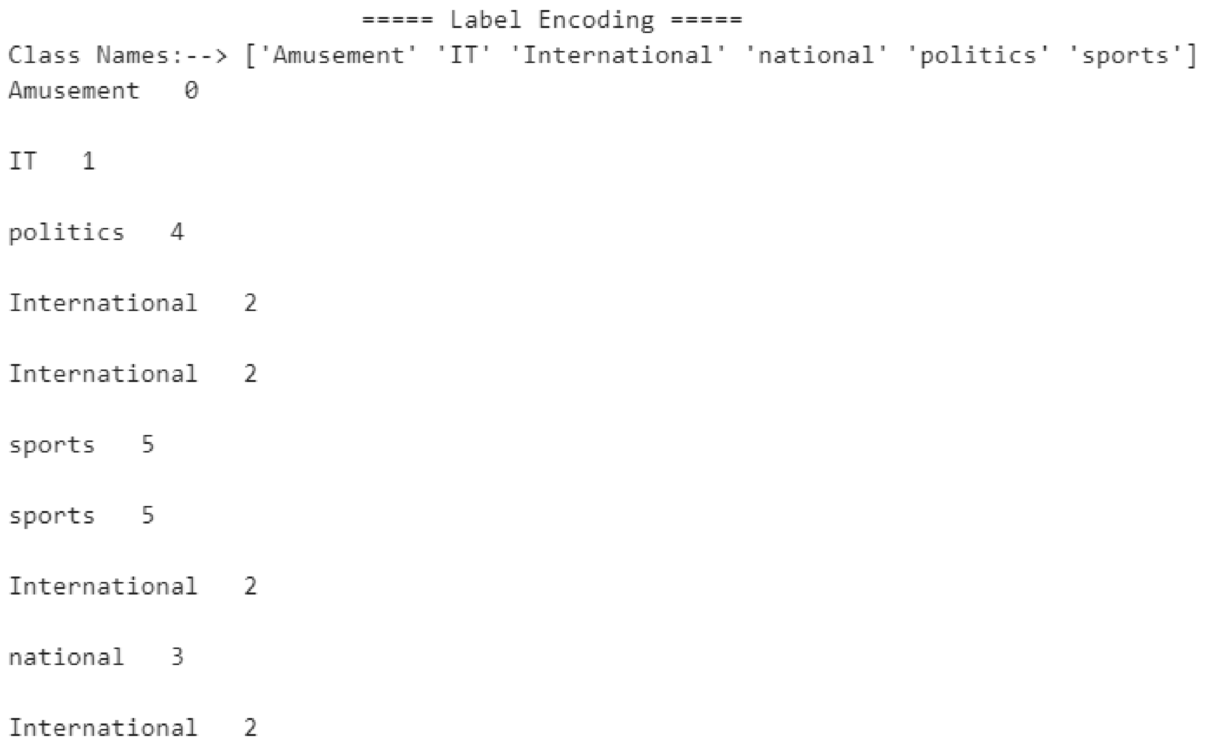
The text data are represented by a encoded sequence where the sequences are the vector of index number of the contains words in each headlines. The categories are also encoded into numeric values.

After preparing the headlines-



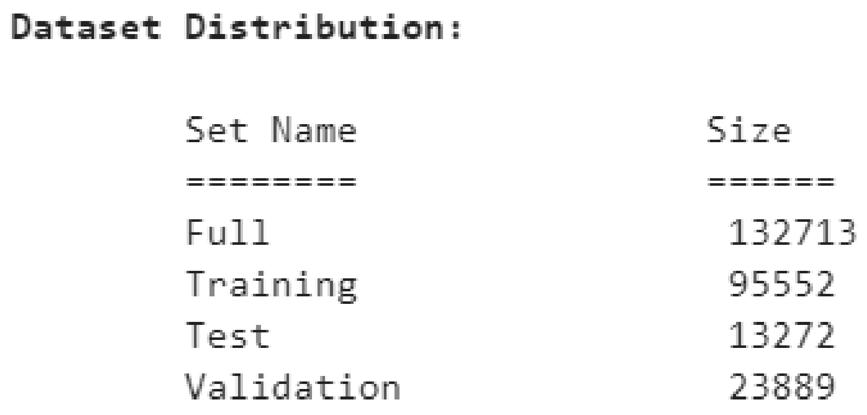
```
                    ====== Encoded Sequences ======

 অসুস্থ হয়ে ফের হাসপাতালে দিলীপ কুমার
  [477, 350, 2638, 1194]

                    ====== Paded Sequences ======
  অসুস্থ হয়ে ফের হাসপাতালে দিলীপ কুমার
  [ 477  350 2638 1194    0    0    0    0    0    0    0    0]
```

And labels it looks as -



```
                        ===== Label Encoding =====
Class Names:--> ['Amusement' 'IT' 'International' 'national' 'politics' 'sports']
Amusement   0

IT   1

politics   4

International   2

International   2

sports   5

sports   5

International   2

national   3

International   2
```

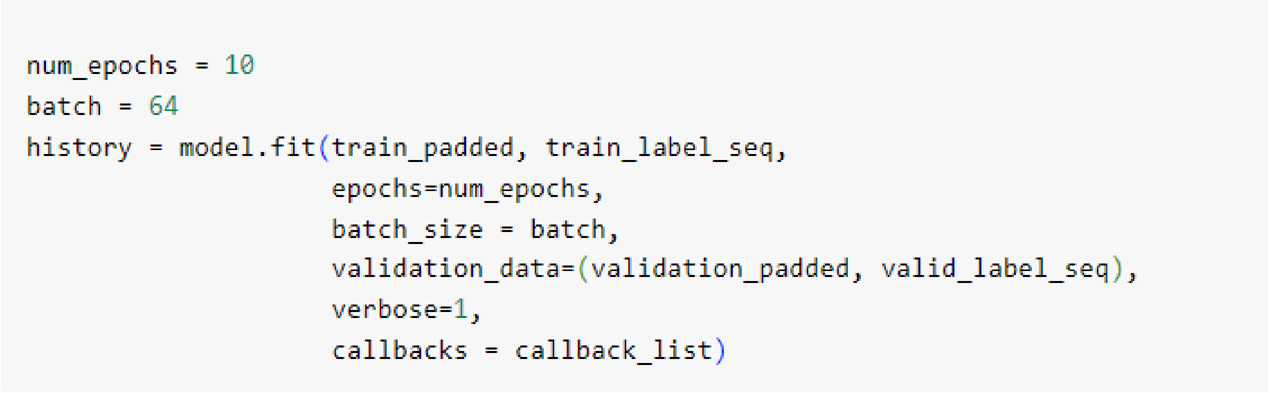
For Model Evaluation the encoded headlines are splitted into Train-Test-Validation Set. The distribution has -



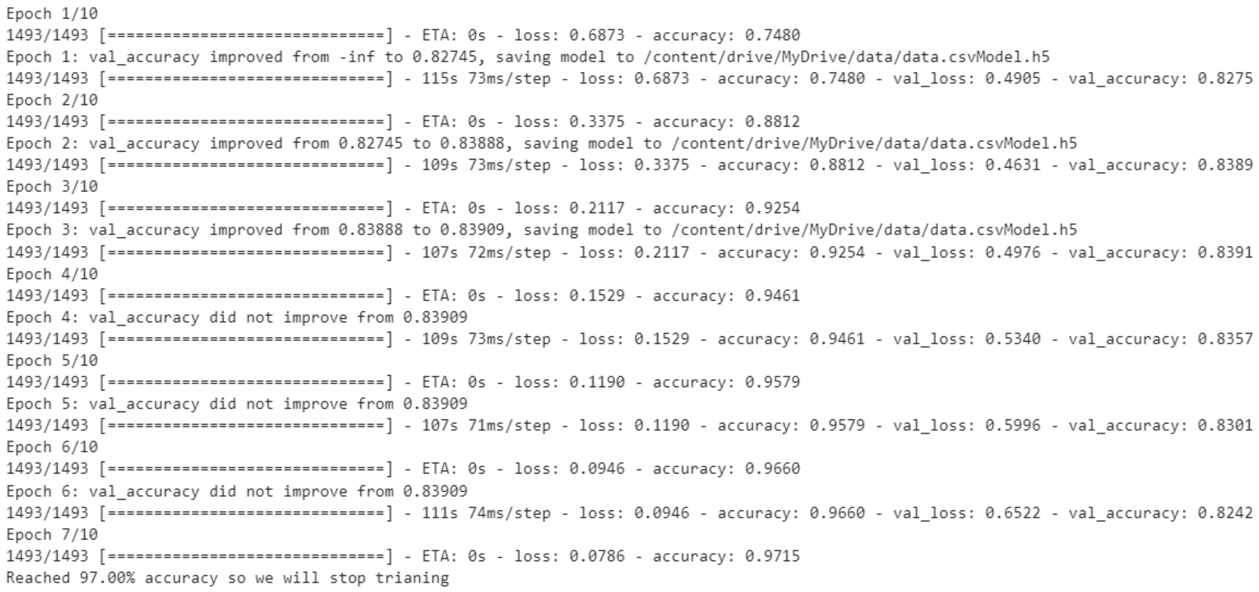
```
Dataset Distribution:

          Set Name                Size
          ========                ======
          Full                     132713
          Training                 95552
          Test                     13272
          Validation               23889
```

5.5 Train Model with Training Dataset

Our model are consist of 5 layers. There are one output layer with 6 category of possible output. In the training, the model achieved 97% accuracy on training data. Though, the dataset is un-balanced.



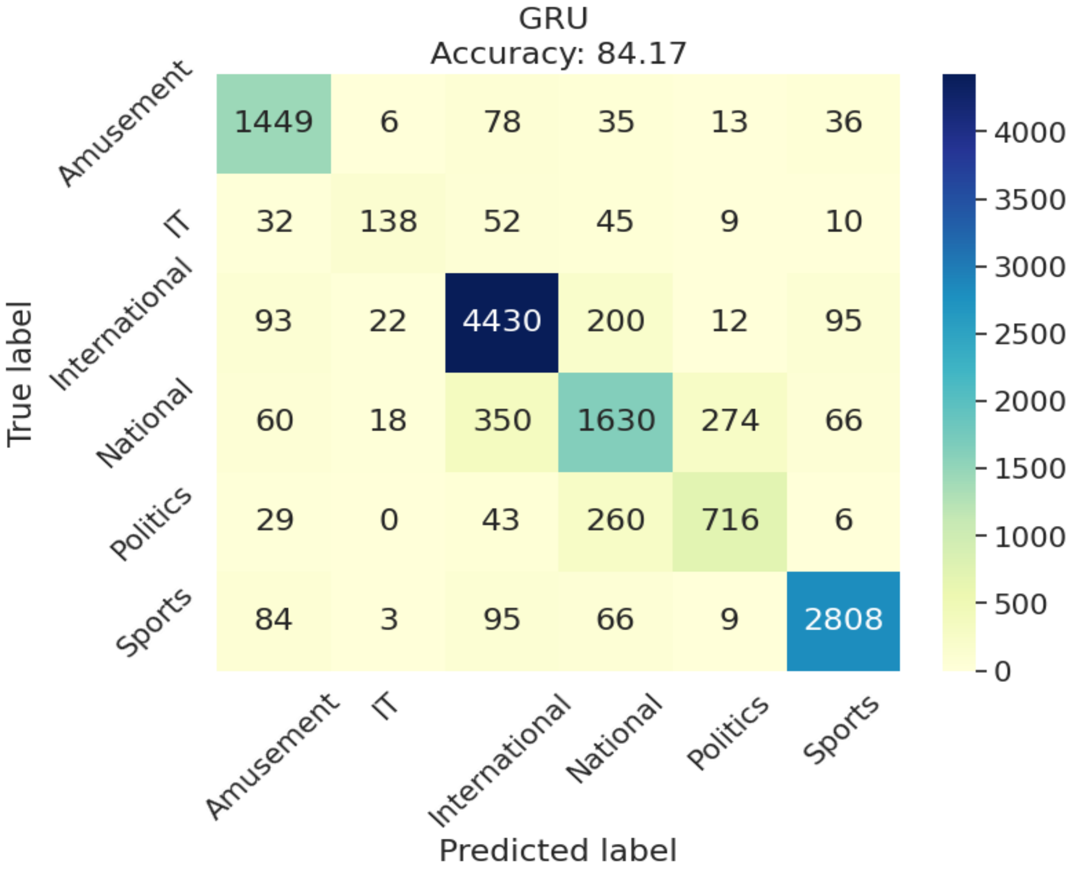
```
num_epochs = 10
batch = 64
history = model.fit(train_padded, train_label_seq,
                    epochs=num_epochs,
                    batch_size = batch,
                    validation_data=(validation_padded, valid_label_seq),
                    verbose=1,
                    callbacks = callback_list)
```



```
Epoch 1/10
1493/1493 [==============================] - ETA: 0s - loss: 0.6873 - accuracy: 0.7480
Epoch 1: val_accuracy improved from -inf to 0.82745, saving model to /content/drive/MyDrive/data/data.csvModel.h5
1493/1493 [==============================] - 115s 73ms/step - loss: 0.6873 - accuracy: 0.7480 - val_loss: 0.4905 - val_accuracy: 0.8275
Epoch 2/10
1493/1493 [==============================] - ETA: 0s - loss: 0.3375 - accuracy: 0.8812
Epoch 2: val_accuracy improved from 0.82745 to 0.83888, saving model to /content/drive/MyDrive/data/data.csvModel.h5
1493/1493 [==============================] - 109s 73ms/step - loss: 0.3375 - accuracy: 0.8812 - val_loss: 0.4631 - val_accuracy: 0.8389
Epoch 3/10
1493/1493 [==============================] - ETA: 0s - loss: 0.2117 - accuracy: 0.9254
Epoch 3: val_accuracy improved from 0.83888 to 0.83909, saving model to /content/drive/MyDrive/data/data.csvModel.h5
1493/1493 [==============================] - 107s 72ms/step - loss: 0.2117 - accuracy: 0.9254 - val_loss: 0.4976 - val_accuracy: 0.8391
Epoch 4/10
1493/1493 [==============================] - ETA: 0s - loss: 0.1529 - accuracy: 0.9461
Epoch 4: val_accuracy did not improve from 0.83909
1493/1493 [==============================] - 109s 73ms/step - loss: 0.1529 - accuracy: 0.9461 - val_loss: 0.5340 - val_accuracy: 0.8357
Epoch 5/10
1493/1493 [==============================] - ETA: 0s - loss: 0.1190 - accuracy: 0.9579
Epoch 5: val_accuracy did not improve from 0.83909
1493/1493 [==============================] - 107s 71ms/step - loss: 0.1190 - accuracy: 0.9579 - val_loss: 0.5996 - val_accuracy: 0.8301
Epoch 6/10
1493/1493 [==============================] - ETA: 0s - loss: 0.0946 - accuracy: 0.9660
Epoch 6: val_accuracy did not improve from 0.83909
1493/1493 [==============================] - 111s 74ms/step - loss: 0.0946 - accuracy: 0.9660 - val_loss: 0.6522 - val_accuracy: 0.8242
Epoch 7/10
1493/1493 [==============================] - ETA: 0s - loss: 0.0786 - accuracy: 0.9715
Reached 97.00% accuracy so we will stop trianing
```

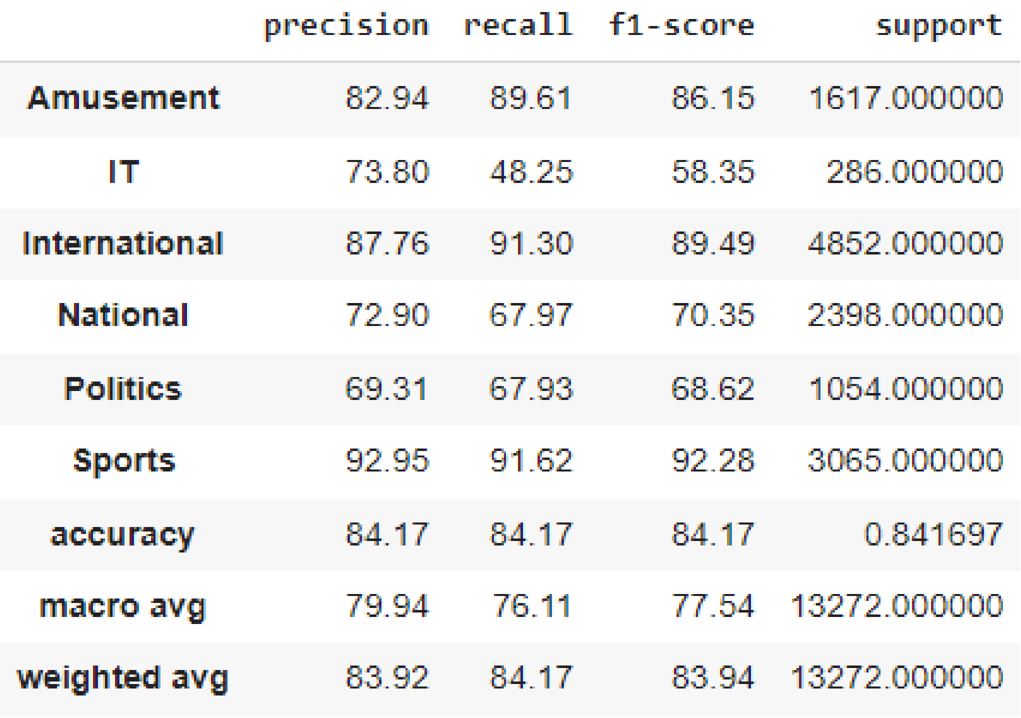
5.6 Model Evaluation

After training, the model was evaluated using the test data and also confusion matrix. Our model performed really well on the test data and the accuracy is 84.17% .Which is very good considered to the dataset and training time.



	Amusement	IT	International	National	Politics	Sports
Amusement	1449	6	78	35	13	36
IT	32	138	52	45	9	10
International	93	22	4430	200	12	95
National	60	18	350	1630	274	66
Politics	29	0	43	260	716	6
Sports	84	3	95	66	9	2808

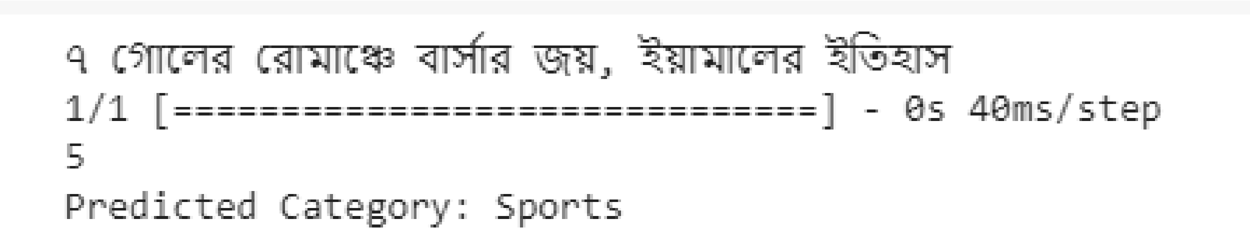
We also evaluated using precision, recall, F-1 score by individual category.



	precision	recall	f1-score	support
Amusement	82.94	89.61	86.15	1617.000000
IT	73.80	48.25	58.35	286.000000
International	87.76	91.30	89.49	4852.000000
National	72.90	67.97	70.35	2398.000000
Politics	69.31	67.93	68.62	1054.000000
Sports	92.95	91.62	92.28	3065.000000
accuracy	84.17	84.17	84.17	0.841697
macro avg	79.94	76.11	77.54	13272.000000
weighted avg	83.92	84.17	83.94	13272.000000

5.7 Output

Our Model can predict News Category correctly. Here is a sample output-



```
৭ গোলের রোমাঞ্চে বার্সার জয়, ইয়ামালের ইতিহাস
1/1 [==============================] - 0s 40ms/step
5
Predicted Category: Sports
```

5.8 Implementation requirement

Google Colab is free platform where anybody can write and execute python programming language using the web browser.it is best for the machine learning, deep learning and data analysis task.

It provide free virtual GPU, which is very good and less time consuming for training a model.

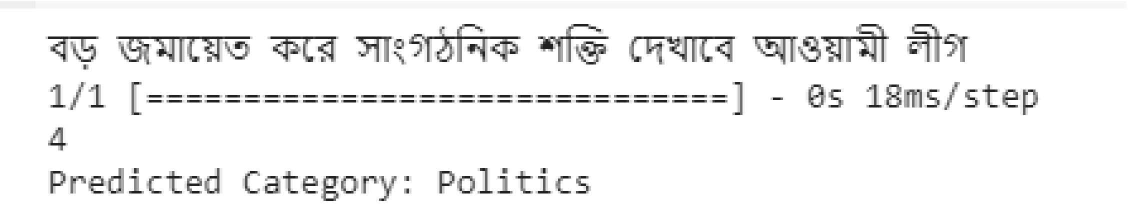
6.Implementation and Experiment

6.1 Description

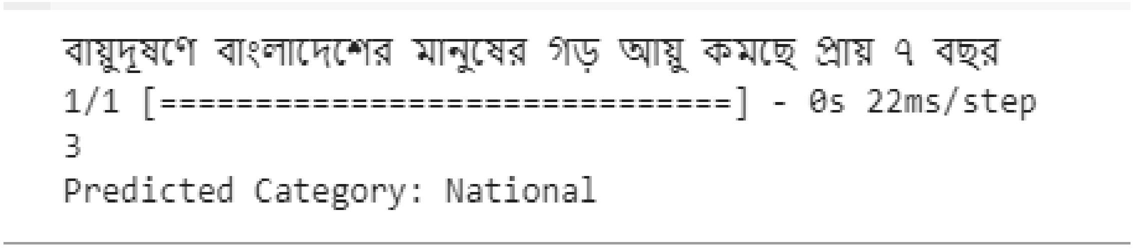
When a new Bangla news headline is provided as input, it undergoes tokenization. This process involves breaking down the headline into individual words or tokens, allowing the model to understand the textual content at a granular level. Sequence Padding: To maintain consistency in the input dimensions, the tokenized sequence is padded with zeros to match the desired length. This step ensures that all input sequences are of the same size, a requirement for feeding data into the neural network. Model Loading: The trained GRU model, which has been saved after the training phase, is loaded into memory. This model embodies the knowledge learned during training and is now ready to make predictions on new input. Prediction Process: The padded and tokenized sequence, now prepared for analysis, is fed into the loaded GRU model. The model processes the sequence through its architecture, capturing the intricate patterns and relationships within the text. Output Categorization: As the sequence is processed, the model generates a probability distribution across the predefined news categories. The category with the highest predicted probability signifies the model's prediction for the input news headline. Display Predicted Category: The predicted category is then interpreted and presented to the user. This efficient process enables users to quickly access the predicted category of the provided news headline, enhancing their engagement with the content

6.2 Experimental Result for Trained Model

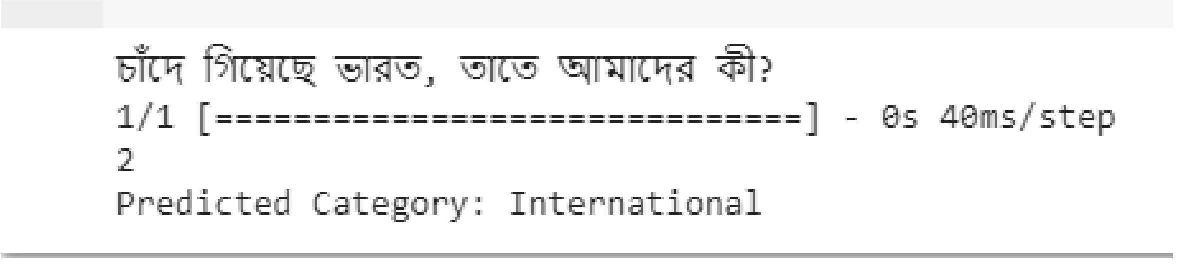
Our Model performed really well on unseen data.



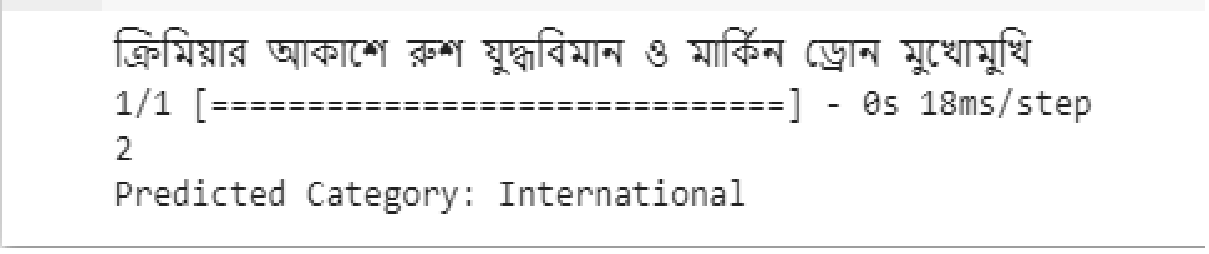
```
বড় জমায়েত করে সাংগঠনিক শক্তি দেখাবে আওয়ামী লীগ
1/1 [==============================] - 0s 18ms/step
4
Predicted Category: Politics
```



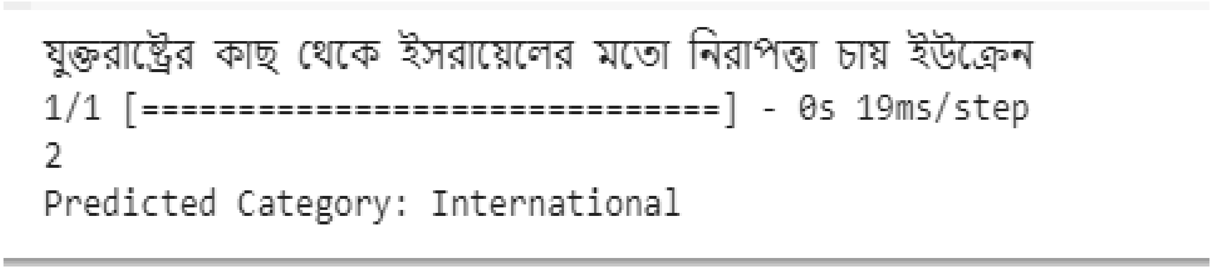
```
বায়ুদূষণে বাংলাদেশের মানুষের গড় আয়ু কমছে প্রায় ৭ বছর
1/1 [==============================] - 0s 22ms/step
3
Predicted Category: National
```



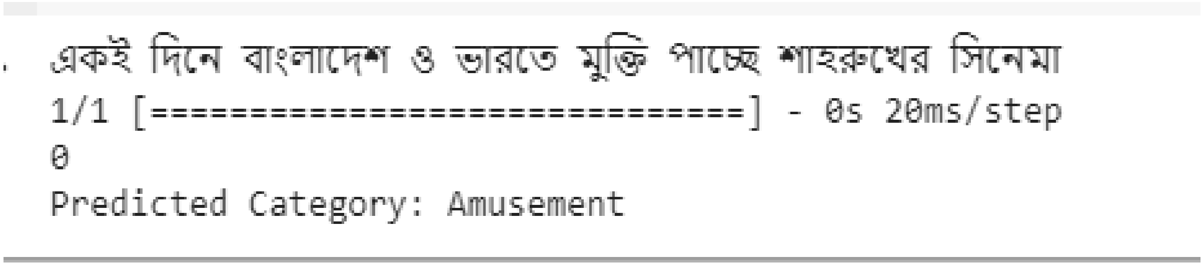
```
চাঁদে গিয়েছে ভারত, তাতে আমাদের কী?
1/1 [==============================] - 0s 40ms/step
2
Predicted Category: International
```



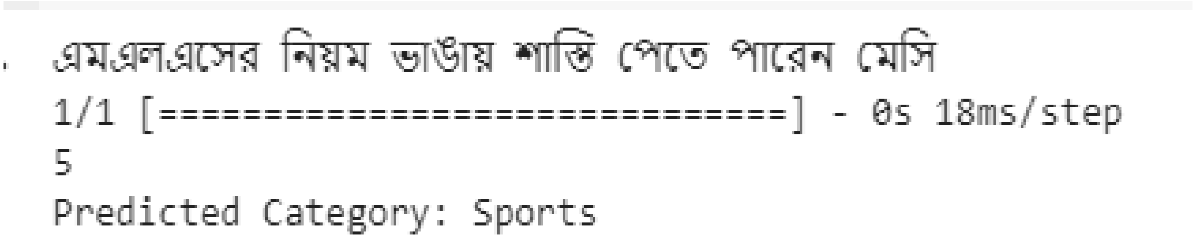
```
ক্রিমিয়ার আকাশে রুশ যুদ্ধবিমান ও মার্কিন ড্রোন মুখোমুখি
1/1 [==============================] - 0s 18ms/step
2
Predicted Category: International
```



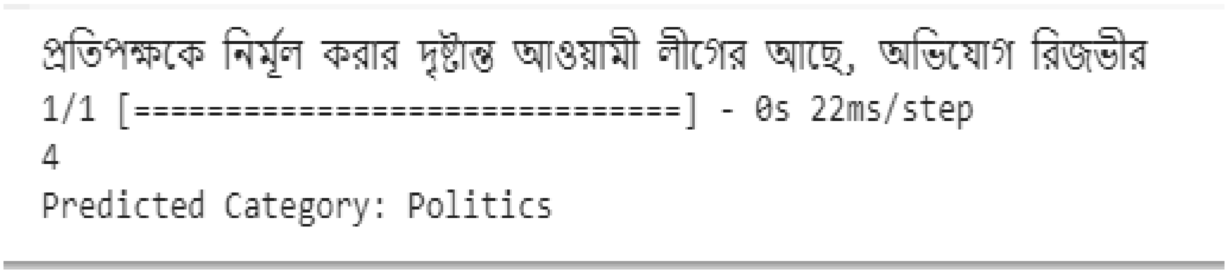
```
যুক্তরাষ্ট্রের কাছ থেকে ইসরায়েলের মতো নিরাপত্তা চায় ইউক্রেন
1/1 [==============================] - 0s 19ms/step
2
Predicted Category: International
```



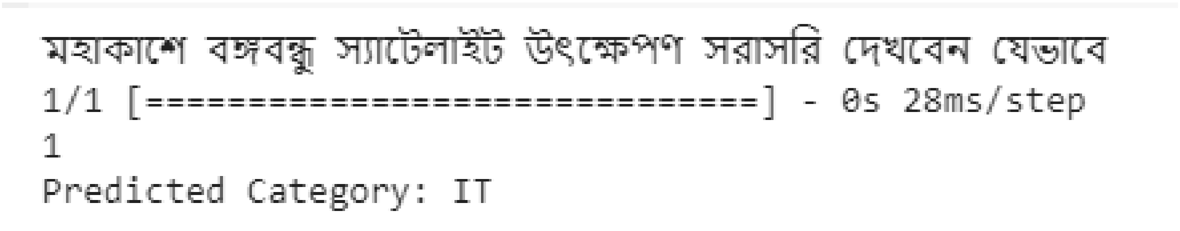
```
একই দিনে বাংলাদেশ ও ভারতে মুক্তি পাচ্ছে শাহরুখের সিনেমা
1/1 [==============================] - 0s 20ms/step
0
Predicted Category: Amusement
```



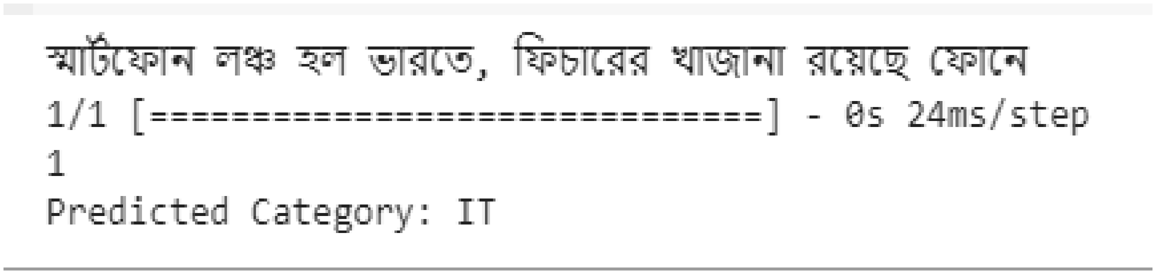
```
এমএলএসের নিয়ম ভাঙায় শাস্তি পেতে পারেন মেসি
1/1 [==============================] - 0s 18ms/step
5
Predicted Category: Sports
```



```
প্রতিপক্ষকে নির্মূল করার দৃষ্টান্ত আওয়ামী লীগের আছে, অভিযোগ রিজভীর
1/1 [==============================] - 0s 22ms/step
4
Predicted Category: Politics
```



```
মহাকাশে বঙ্গবন্ধু স্যাটেলাইট উৎক্ষেপণ সরাসরি দেখবেন যেভাবে
1/1 [==============================] - 0s 28ms/step
1
Predicted Category: IT
```



```
স্মার্টফোন লঞ্চ হল ভারতে, ফিচারের খাজানা রয়েছে ফোনে
1/1 [==============================] - 0s 24ms/step
1
Predicted Category: IT
```

7.Conclusion

CONCLUSION

In conclusion, the culmination of our efforts in this project has yielded an impressive accuracy rate of 84% on an unbalanced dataset. This accomplishment is a testament to the effectiveness of the simple recurrent neural network (RNN) architecture employed for the intricate task of Bengali News Headline Categorization. Our achievement in attaining this notable accuracy underscores the potential of neural networks in deciphering the subtle patterns and inherent complexities present within the Bengali news headlines. By harnessing the power of the RNN, we have successfully navigated the challenges posed by the unbalanced nature of the dataset and achieved a commendable level of precision in classification. However, our pursuit of excellence does not conclude here. There are avenues for further refinement and enhancement. One compelling trajectory is the avenue of hyper-parameter tuning, where the fine-tuning of model parameters could potentially yield even higher levels of accuracy. Additionally, the deployment of more sophisticated network architectures, such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks, holds promise for unlocking even deeper insights from the data. Furthermore, the significance of the dataset size cannot be understated. The potency of neural networks is often magnified with larger datasets, and the acquisition and incorporation of a more extensive Bengali news headline dataset could serve as a catalyst for the model's continuous improvement. In sum, our present accomplishment of an 84% accuracy rate on an unbalanced dataset is an encouraging stepping stone in the realm of Bengali News Headline Categorization. It stands as a testament to the convergence of computational power, linguistic understanding, and advanced neural networks. As we look to the horizon, the potential for elevating accuracy, understanding, and impact beckons, as we strive for more refined, efficient, and accurate models in this vital domain.

7.1 Limitation

There are many limitation we have faced in this project.

- Limited Training Data: Compared to English, the availability of large, high-quality labeled datasets for Bangla news classification is limited. This can lead to challenges in training accurate and robust models.
- Lack of Advanced NLP Tools: While English benefits from a wide range of NLP libraries and tools, the Bangla language has fewer resources, making it challenging to preprocess, tokenize, and analyze text effectively.
- Morphological Complexity: Bangla has a complex morphology, with words often having multiple forms due to tense, gender, and other linguistic features. This complexity can complicate tokenization and feature extraction.
- Domain-Specific Vocabulary: News headlines often use domain-specific terminology, acronyms, and slang, making it important to build a model that understands this domain-specific language.
- Variability in Headline Styles: Bangla news headlines can vary significantly in style, tone, and vocabulary across different news sources. This diversity can pose difficulties for model generalization.
- Contextual Ambiguity: Like any language, Bangla can have ambiguous phrases that require context to interpret correctly. Models might struggle with these nuances, leading to misclassifications.

- Sarcasm and Irony: News headlines sometimes use sarcasm, irony, or humor to convey meaning. Capturing such subtleties is challenging for machine learning models.
- Rare and Specific Categories: Some news categories might be rare or highly specific. If training data is limited for these categories, models might struggle to predict them accurately.
- Lack of Pre-trained Models: While pre-trained models like BERT have been beneficial for English, they might not be readily available or effective for Bangla due to the language's unique characteristics.
- Also it was challenging to train a model on unbalanced dataset.

7.2 Future Implementation

This project in the developed section. There are following possible future works:

- Can create a api to use in other web application.
- Can be used and trained for the sentiment analysis.
- Transfer learning can be done for updated training.
- Ensemble Approaches: Combining predictions from multiple models or using an ensemble of models with different architectures could enhance classification accuracy.
- Multilingual Models: Multilingual models that support Bangla alongside other languages can improve generalization. Such models learn from multiple languages and can be fine-tuned for better accuracy on Bangla tasks.
- Pre-trained Models: With the rise of transformer-based models like BERT, GPT, and their language-specific versions, pre-trained models for Bangla could emerge. These models could provide a strong foundation for various NLP tasks, including news headline classification.
- A User Interface can be created.

REFERENCES

[1] Amin, Ruhul, Nabila Sabrin Sworna, and Nahid Hossain. "Multiclass classification for bangla news tags with parallel cnn using word level data augmentation." 2020 IEEE Region 10 Symposium (TENSYMP). IEEE, 2020.

[2] Ahmad, Istiak, Fahad AlQurashi, and Rashid Mehmood. "Machine and Deep Learning Methods with Manual and Automatic Labelling for News Classification in Bangla Language." arXiv preprint arXiv:2210.10903 (2022).

[3] Salehin, Kamrus, et al. "A comparative study of different text classification approaches for bangla news classification." 2021 24th International Conference on Computer and Information Technology (ICCIT). IEEE, 2021.

[4] Rahman, Md Mahbubur, Md Akib Zabed Khan, and Al Amin Biswas. "Bangla news classification using graph convolutional networks." 2021 International Conference on Computer Communication and Informatics (ICCCI). IEEE, 2021.

[5] Mohiuddin, Ettilla, and Abdul Matin. "Multilevel Categorization of Bengali News Headlines using Bidirectional Gated Recurrent Unit." 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI). IEEE, 2021.

[6] Yeasmin, Sharmin, et al. "Multi-category bangla news classification using machine learning classifiers and multi-layer dense neural network." International Journal of Advanced Computer Science and Applications 12.5 (2021).

[7] Hossain, Mohammad Rabib, Soikot Sarkar, and Moqsadur Rahman. "Different machine learning based approaches of baseline and deep learning models for bengali news categorization." International Journal of Computer Applications 975 (2020): 8887.

[8] Alam, Samrat, Md Afnan Ul Haque, and Ashiqur Rahman. "Bengali text categorization based on deep hybrid CNN–LSTM network with word embedding." *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*. IEEE, 2022.

[9] Rahman, Saifur, and Partha Chakraborty. "Bangla document classification using deep recurrent neural network with BiLSTM." *Proceedings of International Conference on Machine Intelligence and Data Science Applications: MIDAS 2020*. Singapore: Springer Singapore, 2021.

[10] Akanda, Wahiduzzaman, and Ashraf Uddin. "Multi-label bengali article classification using ml-knn algorithm and neural network." *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*. IEEE, 2021.

➢ Data Set Link:

https://www.kaggle.com/datasets/ishtyaquemikrani/bangladesh-news-headlines?fbclid=IwAR3xV9233STSBWJw4tPjLctUZmVhDhvpnQtsfbCqOfUyRuHr9JojGa2DrqI