Bachelor of Science in Computer Science & Engineering

Classification of Bangla News Articles by using Multilayer Perception (MLP)

by

Fatima Jahara

ID: 1504105

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

April, 2021

Classification of Bangla News Articles by using Multilayer Perception (MLP)

Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Fatima Jahara
ID: 1504105

Supervised by

Dr. Mohammed Moshiul Hoque
Professor
Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)
Chattogram-4349, Bangladesh.

The thesis titled “**Classification of Bangla News Articles by using Multilayer Perception (MLP)”** submitted by ID: 1504105, Session 2019-2020 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

Chairman

Dr. Mohammed Moshiul Hoque
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (Ex-Officio)

Dr. Asaduzzaman
Professor & Head
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (External)

Dr. Mohammad Shamsul Arefin
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and that neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Signature of the candidate

Date:

Acknowledgements

The success and outcome of this thesis required a lot of guidance and assistance from several people without whom this thesis would not have been possible and I am extremely privileged to have got this throughout the completion of my thesis. This has been an enriching experience, professionally and personally. All that I have been able to achieve is only due to such guidance and assistance. First and foremost, I would like to express my deep gratitude to my thesis supervisor, Dr. Mohammed Moshiul Hoque, Professor, Department of Computer Science Engineering, Chittagong University of Engineering Technology (CUET) for his continuous guidance, inspiration, and endless support that helped me throughout the preparation of this thesis work. I am thankful for infinite patience, understanding, willingness, and constructive suggestions throughout the entire period. I owe my gratitude to Omar Sharif, Lecturer, Department of Computer Science and Engineering, Chittagong University of Engineering and Technology (CUET) for taking a keen interest in my work and guiding me by providing the necessary expertise whenever required. I am grateful to him for his constant encouragement, and his stimulating ideas. I am incredibly grateful to Prof. Dr. Asaduzzaman, Head, Department of Computer Science Engineering, Chittagong University of Engineering Technology (CUET) for his outstanding support throughout my undergraduate education. I am thankful for the constant encouragement, support, and guidance from all Teaching staff of the Department of Computer Science Engineering, CUET. I would like to acknowledge the constant help and support of the members of the CUET NLP lab who made this work possible through their crucial assistance. I would also like to thank my defense committee members, for providing me with their valuable advice, and feedback.

Abstract

Text classification has growing interest among NLP experts due to the enormous availability of people's textual contents and its emergence on various Web 2.0 applications/services. News article classification is a field of text classification where news documents are analyzed and classified into predefined categories based on their context using NLP. Textual news article classification in the Bengali is also gradually being considered as an important task for online news portals and newsfeed websites to make it easier for the readers to find articles of their preferred categories and for authors to assign articles to the best-suited category. Although the classification of news articles has achieved enormous progressed in high-resourced languages, it is in a preliminary stage till to date in the realm of resourced-constraint languages like Bengali. It is a very challenging task to develop an automatic news articles classification system in Bengali due to its unavailability of necessary resources, shortages of NLP tools and deficiency of benchmark corpora. Automatic classification of huge amount of news articles into categories and subcategories could help in sorting and organizing these textual data. This thesis proposes an Multilayer Perceptron or MLP-based automatic news classification system to classify Bengali news primarily into four categories. The proposed system also can classify four main news classes into 29 subcategories. In this work, we develop a corpus of containing 76343 news articles with 523403 unique words. This thesis explores several word embedding techniques such as Word2Vec, FastText and Keras to find the appropriate features for news article classification task. To investigate the performance of Bengali news classification task, several machine learning-based models (such as LR, DT, RF, k-NN, NB, and SVM) including the proposed MLP-based technique is investigated on developed corpus with both news articles and headlines. The comparative analysis revealed that MLP-based method with Keras embedding layer outperforms the other techniques with an accuracy of 98.18% (for news classification) and 90.63% (for sub-news classification) respectively.

Keywords— **Natural language processing, Text classification, News article classification, Word embedding, Machine learning, Corpus, Bengali language processing and Evaluation**

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANN Artificial Neural Network. 15

Bi-LSTM Bidirectional LSTM. 15

CNN Convolutional Neural Network. 14

HAN Hierarchical Attention Network. 15

LSTM Long Short Term Memory. 15

NLP Natural Language Processing. 5

RNN Recurrent Neural Network. 15

SVM Support Vector Machine. 14

Chapter 1

Introduction

1.1 Introduction

In the era of the Internet, an immense amount of news article is published daily, both online and offline. With the rapid increase of online news sources and the availability of the Internet, people are now more into reading daily news from online news portals. Due to the diversity of information, news classification is not confined only to categories but also to subcategories. Bangla is spoken by about 245 million people of Bangladesh and two states of India, with being the 7th most widely spoken language [1], resulting in an enormous amount of online news articles and text documents written in Bangla. But the articles are not organized or sorted, which makes them difficult to deal with.

This thesis proposes a news classification system for Bengali articles that can classify news articles and headlines into both categories and subcategories. The overview of the Bengali news article classification framework is explained in this Chapter. This Chapter also explains the difficulties, applications, motivation, and contribution of the thesis. Finally, the organization of the thesis is presented at the end of this Chapter.

1.2 News Text Classification

Define the News Text Classification with an example. News Text Classification is the process of classifying or tagging texts and documents collected from news articles into their corresponding categories which are defined previously. News classifiers can automatically analyze text and assign categories to them by using Natural Language Processing (NLP). Let, ND = nd_1, nd_2, nd_3, ..., nd_n is a set of news documents and C = c_1, c_2, c_3, ..., c_m is the set of predefined categories, then news text classification is the assignment of a news document nd_i ϵ ND to a category c_j ϵ C. Each category as a predefined context based on which the documents are assigned to the categories. If SC_j = sc_{j1}, sc_{j2}, sc_{j3}, ..., sc_{jp} is the set of subcategories under category the c_j then subcategorization assigns the news document nd_i ϵ ND to a subcategory sc_{jk} ϵ SC_j. Figure 1.1 depicts the News Text Classification system.

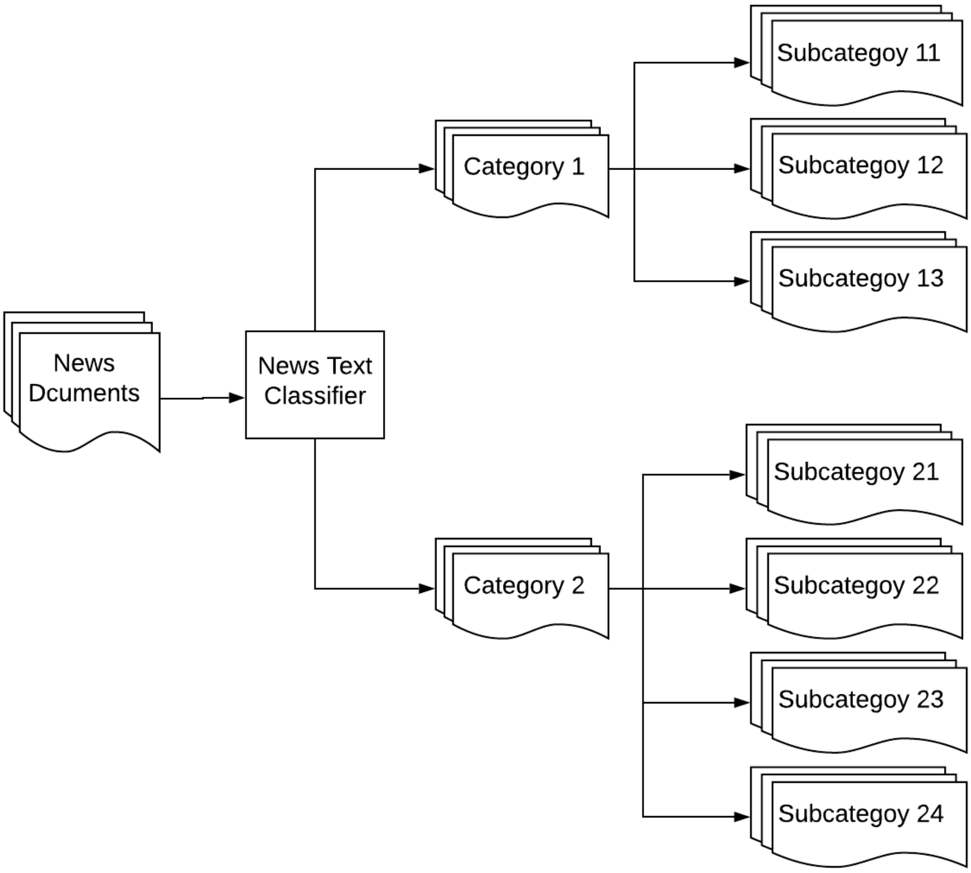


Figure 1.1: News Text Classification System

1.3 General Framework of News Classification

In the field of text classification, news classification is widely considered and is an active research field. With more and more articles publicized daily, it has become almost impracticable for a group of editorials to categorize this massive amount of news articles by reading each one of them. As people have diverse choices, to connect them with their preferred content, we need to know what the editorial is about and then classify them accordingly. Automated news classification systems can deal with these huge amounts of online news articles published daily. The overall news classification framework for classifying news documents is given in Figure 1.2. include the feature extraction with training & testing module in the Fig.Done

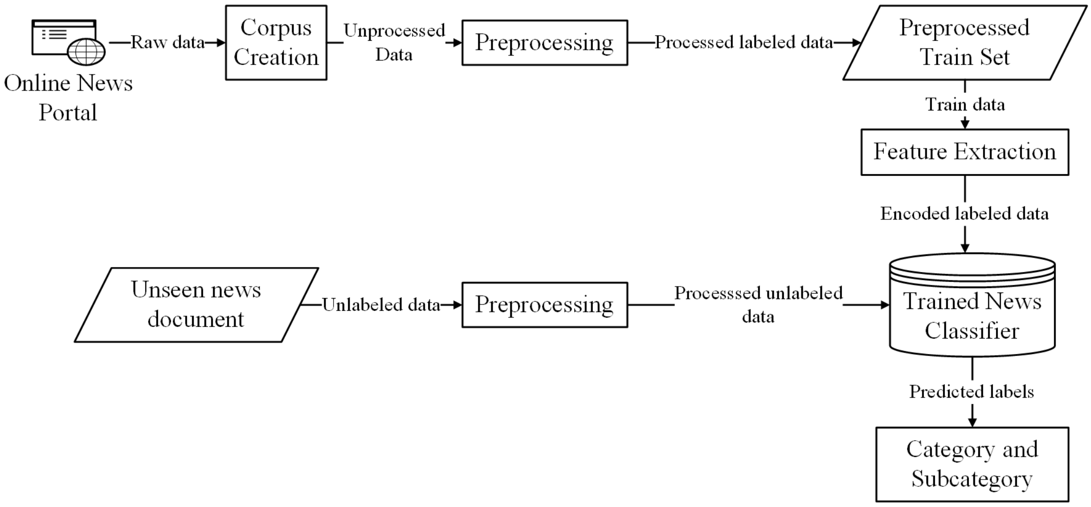


Figure 1.2: Block Diagram of News Classification Framework

The steps followed in this framework are:

- Raw news data are collected for creating a corpus.
- The raw news data are preprocessed and a train set is generated.
- Features are extracted from the processed train data from which a trained news classifier model is generated.
- The unlabeled unseen news documents are preprocessed.
- The processed unlabeled documents go through the classifier model that predicts the label of the document and gives the category and subcategory it belongs to.

1.4 Difficulties

Classification of news articles in Bangla is strenuous due to the lack of standard annotated corpus and complex morphological order in Bangla language. Again topics in Bangla articles are susceptible to multiple labels, making it difficult to designate a news article into a specific class, specifically sub-categories. Preprocessing of these data for training the classification model is also challenging and time-consuming due the unstructured and noisy nature of text. We also need to process a enormous amount of Bangla text data which requires high computational power. One of the hurdles with news article classification is that they are not equally distributable among different classes since some contents make less news compared to others. Sub-classification makes it more undistributable, resulting in an imbalanced dataset. An imbalanced dataset impacts the accuracy of the classifier model. Therefore, handling the imbalanced dataset itself is an important aspect, particularly in the sub-categorization of news articles. Keeping in mind this scenario, we tried working with a real-world imbalanced dataset. Although the dataset is imbalanced, we managed to acquire an accuracy of 98% while categorizing it into 4 main classes and an accuracy of 90% while sub-categorizing it into 29 sub-classes.

1.5 Applications

Rewrite it with point-by-point.Done

Classification is one of the most essential tasks of Text Analysis. News classification is the activity of labeling natural language texts with relevant tags from a predefined set of labels. With the emerging use of the Internet, online news documents are being generated and used more frequently which requires to be handled carefully. Some of the applications of news classification are:

- Thousands of Bangla news sites are constantly providing updated news articles daily. Most of these are unorganized, making it hard to use for analysis, which has created a need to organize Bangla documents intelligently so that users can easily identify required or related documents. Classification of this enormous amount of news articles can help in creating an organized standard resource of news articles for Bangla.

- Automated Bangla news tagging can help in making the user experience better and also in web searching using tags.Online news portals need to sort news articles to make it easier for readers to search for articles in their preferred category. This can help readers to find their preferred articles effortlessly.
- While uploading news articles authors are tasked to select the most relevant category so that the new article can be grouped with similar articles. Since selecting the relevant category should not depend on the authors' opinion but a standard maintained by the website, it can be hard for authors to understand the website standard. In this situation, automated news categorization based on the online portals can save the author from this hardship.

1.6 Motivation

Every day thousands of online news articles in Bangla are being read by millions of users across Bangladesh and several parts of India, which leads to the processing of a massive amount of online textual data. Classifying these can be time-consuming as none have any automatic monitoring or retraining pipeline leading to huge maintenance consuming both time and human labor- something that can easily be circumvented using Natural Language Processing (NLP) along with different machine learning models. There are different types of machine learning techniques, among which deep learning models are more reliable for the task of categorizing these vast amounts of news articles because as the scale of data increase performance of deep learning algorithms increase whereas machine learning algorithms do not perform well.

Category and sub-category classification for news articles are text classification problems where the goal is to assign pre-defined labels to a news article based on its content. There exist well-established research on News article classification in various languages, especially in English. Bengali being the seventh most spoken language by the number of speakers in the world and with the increase in Bangla news articles on the web, the demand for a persuasive and robust automated classification system becomes a must. The research attempts at addressing this problem for Bangla are limited to categorization only. Besides and unlike previous research works, we have employed Multilayer Perceptron model with different feature extraction methods for investigating both categorization as well as sub-categorization of Bangla News Articles and provide comparative results by tuning the model using various hyper-parameters. We have also tried classifying into both categories and sub-categories based on the heading of the articles.

There are several datasets in Bangla for the classification of news articles but none for sub-classification which motivated us to create a new dataset for sub-classification of news articles in Bangla.

1.7 Contribution of the thesis

The main objective of this thesis is to develop a news classification system for categorizing and sub-categorizing Bengali news articles using Multilayer Perceptron technique. The significant contributions illustrates in the following:

- Develop a corpus containing 76343 Bengali news articles with xxx 523403 unique words and xxx 29 classes.
- Investigate various word embedding techniques including mention methods Keras Embedding Layer, Word2Vec and FastText with hyperparameters tuning for Bengali news classification.
- Propose a MLP-based method to classify Bengali news articles into 4 news classes and 29 sub news classes.
- Investigate and compare the performance of the proposed model with other ML baselines and existing techniques.

1.8 Thesis Organization

The overall thesis report is distributed into five chapters. The thesis is organized as follows:

- Chapter 1 contains an introductory explanation of the overall project along with the difficulties the project has faced, the motivations regarding the work, and the contributions we have made in this project.
- Chapter 2 contains a brief discussion of the previous works in the field of text classification that have already been implemented along with their limitations and how our proposed system solves them. It also gives an idea about some concepts related to our system and the challenges we faced while implementing our system.
- Chapter 3 presents our proposed system along with necessary diagrams, algorithms, and tables. The implementation requirements are also mentioned.
- Chapter 4 presents the implementation requirements and impact of our proposed method.
- Chapter 5 presents the overall experimental results along with the setup and procedure. It also gives a view of the impact of our proposed work and gives a comparison of our model with a few other models.
- Chapter 6 gives the conclusion with a summary of our system along with the plans we have for further development of our system.

1.9 Conclusion

In this chapter, we have given a brief introduction to our overall system. The design overview of our system along with the difficulties that follow our work is discussed. The places where our system can be applied are mentioned along with the motivation behind developing the system. The of our work is briefly mentioned in this chapter.

Chapter 2

Literature Review

2.1 Introduction

They are several existing research works on news article classification in the field of text classification in different languages, most of which are based on rule-based and classical machine learning techniques. With the popularity of deep learning algorithms increasing steadily, research on news article classification using deep learning techniques is increasing. Chapter 2 explains the a few basic terminology and techniques. Moreover, this Chapter discuss the various news classification techniques that is closely related to the proposed work. Few implementation challenges also highlights at the end of this Chapter.

2.2 Basic Definitions

There are several concepts related to text and news classification system. Basic definitions of the related concepts are explained in this section. Write all definitions with proper citation.Done

2.2.1 Text Classification

Text Classification is the task of classifying a document under a predefined category [2]. Let, D = d_1, d_2, d_3, ..., d_n is a set of documents and C = c_1, c_2, c_3, ..., c_m is the set of predefined categories, then text classification is the assignment of a document d_i ϵ D to a category c_j ϵ C. Each category as a predefined context based on which the documents are assigned to the categories.

Text classification is a wide field that includes news classification, fake news detection, spam filtering, suspicious text detection, emotion detection, and many others. Due to an excessive increase in online textual data such as emails, messages, online news articles, comments, etc text classification is becoming more open for research. Text classification is now used in healthcare, crime investigation, depression analysis, and many more fields.

Texts are a form of natural language that requires preprocessing before they can be used to train a text classifier model. Tokenization, stopword removal, text encoding, sequence padding, etc are a few preprocessing tasks that are required to feed the texts to the classifier for text classification.

2.2.2 News Classification

News classification is a field of text classification where news documents are classified into their corresponding predefined labels. These labels can be both categories and subcategories. Documents used for classifying news can be both articles and headlines. Automated news classification is an intelligent classification of text into categories using Natural Language Processing. Automating these tasks using artificial intelligence, simply makes the entire process super-fast and efficient. Machine learning and deep learning techniques with the use of natural language processing make it possible.

News documents are a form of text that contains a rich and informative type of data. These text documents can be used in various tasks such as topic extraction, summarization, category and subcategory classification, etc. But these raw text needs to be converted into numerical values for classifier models to understand. Hence to utilize text data in machine learning these textual data are first needed to be preprocessed and transformed to machine-understandable encoded values. Then these encoded values can be used to build classifiers that can work through supervised learning and unsupervised learning. Supervised classification is done when classification categories are predefined and the text documents are labeled. When the classification categories are unknown unsupervised learning is applied.

2.2.3 Word Embedding

Word embedding is a Natural Language Processing technique that represents words in a text, based on the context and neighboring words. Word embeddings capture both syntactic and semantic information of words and capture relations between them based on context and morphology[3]. Word embeddings represent words with similar meanings with similar representations. They represent words using real-valued vectors that encode the meaning of the words in a way that words closer in the vector space are assumed to have similarities in meanings and dissimilar words are assumed to be far away from each other. Word embeddings can be obtained by mapping words from a vocabulary to vectors using real numbers. They also help by reducing the dimensionality by plotting words from multi-dimensional space to a much lower dimension. The embedding values are parameters learned with the help of gradient descent.

Several well-known word embedding techniques are described below.

Embedding Layer: The embedding layer is a word embedding technique that is learned jointly with deep learning models on different NLP tasks such as language modeling, text classification, etc. Embedding layers perform embedding on the input layer of a neural network model. The embedding layer available in Keras can be initialized with different embedding dimensions in the vector space. The keras embedding layer initializes word embedding using some random values from a uniform distribution and updates them while training the neural network model. They are used at the beginning of a neural network connected with the layers such that they can be updated through backpropagation.

Word2vec: Word2vec is a model that processes text by vectorizing words and create word embeddings. It is a statistical method used for neural network-based training. There are two learning models available with word2vec, namely Continuous Bag-of-Words, or CBOW model, and Continuous Skip-Gram Model. Continuous Skip-Gram Model. Here the tokenized dataset is given as input into the model and outputs a set of feature vectors. We have used Gensim for extracting features using the Word2Vec model. We have used both Continuous Bags of Words (CBOW) and Skip-gram Word2vec models. CBOW learns the probability of a word based on the context surrounding it. The skip-gram model predicts the surrounding context words given a target word.

FastText: FastText is a word embedding technique that is an extension of word2vec. FastText embeddings exploit subword information to construct word embeddings where representations are learned of character n-grams, and words represented as the sum of the n-gram vectors [4]. fastText manages rare words very well. Even if a word is not seen during training, it can be split down into n-grams to get the embeddings. Continuous Bag of Words (CBOW) and Skip-gram fastText are two types of fastText embedding models.

2.2.4 Multilayer Perceptron Model

Multilayer perceptron (MLP) [5] is a feedforward artificial neural network consisting of an input layer and an output layer connected by multiple hidden layers in between. A perceptron is a single neuron model which builds up a multilayer perceptron when stacked in layers. The neurons also known as units in these layers are fully connected in a chronological way (input to output), typically referred to as feedforward: input layer -> hidden layers -> output layer.

The input layer takes in the input values and passes them to the hidden layers. No processing of data occurs in this layer. The neurons in the hidden layers perform the classification of the features with the help of non-linear activation functions and predefined weights and bias. The final output prediction calculated within the hidden layers is passed through the output layer. The layers in MLPs are also known as Dense layers. MLPs employ supervised learning which is carried out through backpropagation. Learning occurs in the MLPs through changes in weights in each neuron by comparing the expected output to the amount of error generated.

A classical multilayer perceptron consists of the following functions:

- A linear function is a function that aggregates the input values in each neuron.

- An activation function also called the sigmoid function which is a non-linear function used on the output of the linear function.
- A loss or cost function that measures the error of the MLP network by comparing the output of the network with the actual output.
-

Multilayer Perceptrons operate in two different steps, namely forward propagation and backward propagation.

1. **Forward propagation:** In forward propagation for the given inputs, also known as features, outputs are calculated using the initial weights and bias which are initialized randomly. At first, the weighted sum of input and bias is calculated using the linear function, which is then used to calculate the output applying the non-linear activation function. These functions are applied to each of the neurons. The predicted output of the activation function at the output layer is then used to compute the loss. The overall process is known as forward propagation.
2. **Backward propagation:** Backpropagation is used for training the model with proper parameters. Different optimizers along with different learning rates are used to minimize the loss calculated in the forward propagation by updating the weights and the bias. This process of minimizing loss is called backpropagation.

Several terms related to Multilayer Perceptron models are given shortly.

Regularization: Neural networks tend to memorize the training data which results in overfitting. Regularization is used in MLPs to avoid overfitting. Regulatory layers like Dropout can be used for the cancelation of a fraction of units while training the models.

Optimization: Optimization is the process used to minimize the loss function. Optimizers tend to reduce loss to an acceptable value so that the model learns to map the inputs to the outputs correctly. Adam optimizer is an optimization technique for gradient descent that maintains a learning rate for updating weights. Learning rates remain constant throughout the training.

Batch size: Batch size is the number of training data that are going to be propagated through the network in one forward or backward pass. Since weights are updated after each propagation batches help in training faster.

Epoch: Epoch defines the number of times a model works on the entire train set. One epoch indicates that the training dataset has gone through a forward and a backward pass while training. An epoch can consist of one or more batches of data.

2.3 Related Work

Works related to news classifications can be described into 2 sectors. Machine Learning based and Deep Learning based. Organize this section as ML based and DL based.

2.3.1 ML based News Classification

There have been many works on news classification based on different aspects and using different model frameworks in different languages. Most of the previous works are based on machine learning techniques. The authors in [6] have used Naïve Bayes, Decision Trees, Random Forest, Support Vector classifiers, and Multi-layer Perceptron for news article text classification and summarization based on both authors and topics. A technique for categorizing online news articles based on semantic similarity and Word-Net has been proposed in [7], where the news articles are categorized using user-annotation based technique by keyword identification and extraction of semantically similar words from Word-Net. In [8], using K-Nearest Neighbor for Indonesian news identification of classes in each multi-label news article is done. There have not been many works on the subcategorization of news articles. The authors in [9] proposed two different algorithms for category classification and topic discovery and classification of Japanese and English news articles where at first keyword extraction is used for category classification and later using both keyword extraction and one-pass clustering topic discovery and classification is completed.

In recent years with the advancement of Natural Language Processing and different machine learning techniques there has been an increasing amount of research on Bangla document and news article classification. Four supervised learning methods Decision Tree, K-Nearest Neighbour (KNN), Naïve Bays (NB), and Support Vector Machine (SVM) were used for categorization of Bangla web documents into 5 different categories in [10] where the authors created a dataset containing 1000 Bangla web documents. In [11], Logistic Regression, Neural Network, Naive Bayes, Random Forest and Adaboost models using Word2Vec and TF-IDF feature extraction methods has been used for classification of Bangla news articles into five different categories. The authors in [12] have used the Naive Bayes classifier to classify Bangla news articles based on news code of IPTC.

2.3.2 DL-based News Classification

Recently there have been many works on news article classification using deep learning techniques. One of which is [13], where the authors have used three different models: SVM, MAXENT, and CNN in two different schemes for the classification of Chinese news into 12 different categories using a dataset of 9563 articles. For both SVM and MAXENT models, the TF-IDF matrix has been used for feature extraction, and for the CNN model an embedding layer has been used. MAXENT model showed the best performance, with an overall accuracy of 93.71%. CNN with Continuous Bag-of-Word(CBOW), CNN with Skip-gram, and CNN without word2vec models are used for real news articles and tweets classification in [14] where CNN classification model with CBOW showed higher performance in news article classification. Z. Lu et al. [15] proposed an efficient approach for Chinese news headline classification into 18 categories based on a multi-representation mixed model with attention and ensemble learning accruing an F1 score of 0.8176.

In recent years with the advancement in deep learning techniques, there have been many works on news classification in Bangla using deep learning techniques. In [16], the authors have fine-tuned multilingual transformer models- multilingual BERT and XLM-RoBERTa, for Bangla text classification in different domains, including news categorization. Three different datasets were used for classification using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) models that encoded the documents at their character level [17]. In [18] different machine learning based approaches for baseline evaluation and Bi-LSTM and CNN for fine-tuned predictions are used for Bengali News categorization into 12 categories. A framework for classifying Bengali documents using deep convolution nets with Word2Vec(skip-gram) is proposed where more than 1 million Bengali text documents is used for categorization into 12 categories [19]. Several traditional machine learning methods and advanced deep learning based supervised methods including MLP, CNN, Simple RNN, LSTM, C-LSTM, Bi-LSTM, HAN, Convolutional HAN were used for both news article and title classification of Bangla news documents into 10 predefined categories in [20]. In [21], the authors proposed a text classification method using bidirectional LSTM with attention mechanism to classify Bangla news articles into 12 different categories based on the news captions. The authors in [22] created a dataset for classification of Bangla news titles into 7 different categories using Parallel CNN. In [23] the autors created a dataset for classifying Bangla news headline into 8 categories using ANN, SVM, LSTM and Bi-LSTM models.

add a paragraph at the last to summarize the weaknesses/limitations of existing news classification in Bengali. Then write how your system solve/address these limitations

Looking at the previous works related to the news classification it can be observed that most of the works are on news article classification into categories only. There has been no work on Bangla news subcategorization.Most of them are based on classifying 5 to 12 categories of Bengali news articles. To address this problem we have developed a MLP based classification system that can detect and classify articles into subcategories, and also outperforms previous works on Bengali document classification.

2.4 Problem Statement

Today the online portals are not only confined to categorical distribution but also into subcategories. Subcategories make news archiving easier for people of interest. Since thousands of news in Bangla are being published daily, the work of sub-categorizing the news articles by the editorial team is becoming tedious. With the increase in Internet usage, more people are being interested in online news. To make online news articles more manageable, automated online news subclassification for Bangla is becoming crucial. To overcome this situation we have developed a news classification system for Bengali language that can classify news into subcategories. We have also built the system for category classification for cases where subclassification is not required.

Recently there have been few works of news headline classification since headlines are easy to process for their shorter length. For this reason, we have developed our system that can classify both articles and headlines.

For the development of our classification system we had to create our own corpus and label them accordingly. This corpus can be used in the future for more research in news subclassification.

2.4.1 Implementation Challenges

For implementation of our news classification system we had to face some challenges. These challenges are listed below:

- There has been no work on news classification into subcategories. With no previous works on subclassification of news articles, no corpus is available for the development of this system. So we had to create a corpus of our own for developing our system. For this reason, we had to collect news articles and headlines from different online news portals that consists of 76343 news articles and headlines. We had to build a web crawler that could crawl news articles from the portals.
- Since our work is based on supervised learning labeled data for training is a must. So we had to label the dataset into 4 categories and 29 subcategories

in a span of around 1 year. We had to build a system that can automatically label the datasets into subcategories based on some predefined keywords. Since supervised learning depends mostly on the correctness of the labels we had to check manually the news articles to ensure that the labels are correctly given.

- In many cases, the labels of the news articles do not match with the keywords. Some keywords like মারা গেছে। (Died) can occur in cases of all the subcategories under Crime and Accident data. The keyword is also seen in Entertainment and Sports subcategories in case someone dies. In these cases, we had to go through the whole article to make sure the labels are correct.

2.5 Conclusion

A detailed literature review has been discussed in this chapter. For convenience, this study was divided into three sub-sections. Different classification techniques along with dataset information were mentioned to get a precise view of the previous works related to our project. The next chapter contains a complete explanation of the proposed methodology for news classification.

Chapter 3

Methodology

3.1 Introduction

In this chapter, we will discuss the methodology of our proposed news classification system. The modules used for the overall development of our system are explained with necessary figures, tables and algorithms.

3.2 Corpus Development

Explain the procedure to dataset development in details with subsections Done

For the research purpose, we have developed a corpus of our own since there is no dataset available on sub-classification of news articles. We have collected almost 187,031 news articles under the CUET NLP lab from which a corpus of 76,343 news articles is created which is then classified into 4 main categories and 29 subcategories. The overall process of corpus creation is discussed in this section.

3.2.1 Data Crawling or Accumulation

The data collection task was automated by web-scraping the online news portals. In total, we have collected 76,359 news articles from 5 renowned news portals, namely: Prothom Alo [24], Daily Nayadiganta [25], Daily Samakal [26], Kaler Kantho [27], and Bhorer Kagoj [28], as shown in Table 3.6. The news articles span from the year 2015 to the year 2020. Five participants from CUET NLP lab worked on accumulating data from these data sources. The whole data accumulation process was done from August 2019 to October 2020. In this time period we have scraped almost 187,031 news articles from which 76,343 news articles are used to create the corpus for news classification system.

The news crawling was based on four main categories- Accident, Crime, Sports, and Entertainment. Title, Author, Date, and Description are the four main sections that were collected. We have used both the article and the headline for our classification purpose.

3.2.2 Data Preprocessing and Cleaning

Data cleaning is the process of preparing data by removing unnecessary data. Data cleaning prevents coarse data from providing inaccurate results. There are several punctuation marks and special symbols in our dataset, such as @, $, %, *, (,), <, >, , , ?, ., /, :, ; etc, which don't bear any significance to the classification process. Again a general Bengali news article may contain Bengali as well as English digits and also some foreign words. These punctuations, digits and foreign words are removed from the dataset due to their insignificance in model classification. Stop words are most common words found in any natural language which carries very little or no significant semantic context in a sentence [29]. We have removed 398 common Bangla stop words from our dataset to focus on the important words that bear meaning in the context of the data.

The raw data contained 15,655,516 total words and 662996 unique words and after preprocessing and cleaning there remained 11,843,411 total words and 523,403 unique words in the news articles. In the news headlines raw data contained 516781 total words and 60,167 unique words and after preprocessing and cleaning there remained 373,860 total words and 61,470 unique words.

3.2.3 Data Annotation

Data that were already classified into categories were further subclassified. The data annotation part was done by five members of the CUET NLP lab who are final year undergraduate students from Computer Science and Engineering backgrounds, under the supervision of Dr. Mohammed Moshiul Hoque, Professor, Department of Computer Science Engineering, Chittagong University of Engineering Technology (CUET) and Omar Sharif, Lecturer, Department of

Computer Science and Engineering, Chittagong University of Engineering and Technology (CUET).

The data annotation task was performed in two steps. First through an automatic keyword matching system and then through manual checking by the annotators. First of all, a set of subcategories was defined for each category, based on different renowned news portals. Then the initial label was chosen using a keyword matching process is described in Algorithm 1. The keywords were extracted based on the subcategories in Prothom Alo news portal. After that, the five annotators checked the labels manually to ensure the quality of the annotation.

Algorithm 1: Sub classification using keyword matching

```
Enter_SourceDirectory
Enter_DestinationDirectory
F ← $List\ of\ files\ in\ SourceDirectory$
K ← $List\ of\ Keywords$
for file ∈ F do
    $T \leftarrow text\ in\ file$                        ▷ Read text in file
    for key ∈ K do
        if key ∈ T then
            $DestinationDirectory \leftarrow file$       ▷ Move file to DestinationDirectory
        end
    end
end
```

3.2.4 Data Statistics

The data statistics of the dataset that we have used for developing our classification system are mentioned with some sample data examples. Our dataset contains a total of 76343 news documents where the total number of words in news articles is 15,655,516 and in headlines is 516,781. Our news articles contain 662,996 unique words and headlines contain 60,167 unique words in total. Table 5.3 highlights the summary of the raw dataset in each category.

The statistics of the dataset into perspective subcategories along with sample data are summarized in Table 3.2, Table 3.3, Table 3.4 and Table **??** for the Accident, Crime, Entertainment and Sports category respectively.

Class	Data	Total Words		Unique Words	
		Article	Headline	Article	Headline
Accident	11841	2,077,659	74,387	100,018	6,834
Crime	11,222	2,998,583	78,160	127,494	10,175
Entertainment	17,568	2,828,029	120,263	166,258	20174
Sports	35,712	7,751,245	243,971	269,226	22,984
Total	**76,343**	**15,655,516**	**516,781**	**662,996**	**60,167**

Table 3.1: Data statistics in each class

Subcategory	Data	Sample Data
Air	51	চট্টগ্রামের লোহাগাড়ায় বিমানবাহিনীর একটি প্রশিক্ষণ উড়োজাহাজ বিধ্বস্ত হয়েছে।
Blast	427	চট্টগ্রামের পাথরঘাটা এলাকায় বিস্ফোরণের ঘটনা ঘটেছে।
Construction	191	ফেনীতে নির্মাণাধীন ভবনের ছাদ থেকে পড়ে এক শিশুর মৃত্যু হয়েছে।
Electricity	959	পিরোজপুরের ভান্ডারিয়া উপজেলায় বিদ্যুৎস্পৃষ্টে দুই শ্রমিক মারা গেছেন।
Fire	1923	কেরানীগঞ্জের প্লাস্টিক কারখানায় আগুনে দগ্ধ আরও এক ব্যক্তি মারা গেছেন।
Rail	975	রাজধানীর নাখালপাড়ায় ট্রেনে কাটা পড়ে নিহত দুজনের পরিচয় মিলেছে।
Road	5425	মাদারীপুরে পৃথক সড়ক দুর্ঘটনায় তিনজন নিহত হয়েছেন।
Water	1648	বুড়িগঙ্গা নদীতে বালুবাহী একটি বাল্কহেড ডুবে ৪ শ্রমিকের মৃত্যু হয়েছে।
Others	242	চট্টগ্রামের বোয়ালখালী উপজেলায় হাতির আক্রমণে একজনের মৃত্যু হয়েছে।

Table 3.2: Data Statistics of each subcategory in the Accident category

3.2.5 Data Distribution

Table 3.6 shows the sourcewise distribution of the news classifcation data.

Figure 3.1a and 3.1b shows the visual distribution of the dataset based on categories and sub-categories.

Source-wise, Training/validation or testing set-wise. You may use any Visualization tools such as WordCloud to visualize the distribution. Most Important:

Subcategory	Data	Sample Data
Corruption & Fraud	1464	আজিমপুর মাতৃসদন ও শিশুস্বাস্থ্য প্রশিক্ষণ প্রতিষ্ঠানে কেনাকাটায় দুর্নীতির প্রমাণ পেয়েছে দুর্নীতি দমন কমিশন (দুদক)।
Drug	850	আনোয়ারায় ১৫০০ পিস ইয়াবাসহ তিনজনকে গ্রেপ্তার করেছে পুলিশ।
Murder	5205	যশোরে দুর্বৃত্তদের ছুরিকাঘাতে এক যুবক খুন হয়েছেন।
Rape & Abuse	1185	রাজধানীতে পৃথক ঘটনায় যাত্রাবাড়ী ও কদমতলীতে দুই তরুণী ধর্ষণের শিকার হয়েছেন।
Suicide	1271	সাতক্ষীরা সদর উপজেলায় গতকাল শুক্রবার এক গৃহবধূ আত্মহত্যা করেছেন।
Theft & Robbery	554	ঝিনাইদহের শৈলকুপায় মন্দির থেকে ৯৩ কেজি ওজনের শিবলিঙ্গ চুরির ঘটনা ঘটেছে।
Trafficking	285	ব্রাজিলে বাংলাদেশি বংশোদ্ভূত এক মানব পাচারকারীকে গ্রেপ্তার করেছে পুলিশ।
Others	408	প্রশাসনের কোনো পদক্ষেপেই কবজায় আসছে না পেঁয়াজের সিন্ডিকেট।

Table 3.3: Data Statistics of each subcategory in the Crime category

Subcategory	Data	Sample Data
Bollywood	10496	বলিউডের এ সময়ের দুই প্রতিভাবান তারকা কার্তিক আরিয়ান ও সারা আলী খান।
Dhallywood	2065	রুপালি পর্দার জনপ্রিয় নায়ক মান্নার মৃত্যুবার্ষিকী আজ।
Hollywood	493	অস্কার হয়তো প্রথমবারের মতো লিওনার্দো ডি ক্যাপ্রিওর দুয়ারে কড়া নাড়ছে।
Music	1604	দেশের হার্ড রক ব্যান্ড ওয়ারফেইজ। এবার প্লেব্যাক করলো ব্যান্ডদলটি।
Television	2171	আগামী শনিবার রাত ৮টায় আরটিভিতে প্রচারিত হবে নাটক ‘একমুঠো জোনাকি’।
Tollywood	693	সামনে এল শুভশ্রী গঙ্গোপাধ্যায়ের পরবর্তী সিনেমা ‘ধর্মযুদ্ধ’-র পোস্টার।
Others	46	মিস ইংল্যান্ড ২০১৯’ হয়েছেন ভারতীয় বংশোদ্ভূত বাঙালি তরুণী ভাষা মুখার্জি।

Table 3.4: Data Statistics of each subcategory in the Entertainment category

Use Kappa/Jacard similarity to measure the quality of the dataset and inter-annotator agreement.

Subcategory	Data	Sample Data
Athletics	92	১৩তম এসএ গেমসে বাংলাদেশ পেয়েছে ১৯টি স্বর্ণ, ৩৩টি রৌপ্য এবং ৯০ টি ব্রোঞ্জ।
Cricket	25768	আজ সন্ধ্যায় সিলেট যাবে বাংলাদেশ জাতীয় ক্রিকেটের ওয়ানডে দল।
Football	9264	জাতীয় ফুটবলে চমক দেখিয়েই চলেছেন সাবিনা খাতুন ও কৃষ্ণা রানী।
Tennis	565	দীর্ঘদিনের বিরতির পর টেনিসের কোর্টে নামছেন সানিয়া মির্জা।
Others	23	বাংলাদেশ ভলিবলের জন্য বিশাল এক সুযোগ। সামনে বাধা শুধু শ্রীলঙ্কা।

Table 3.5: Data Statistics of each subcategory in the Sports category

News Portal	Data
Kaler Kantho	41950
Prothom Alo	20613
Bhorer Kagoj	10795
Daily Nayadiganta	2066
Daily Samakal	919

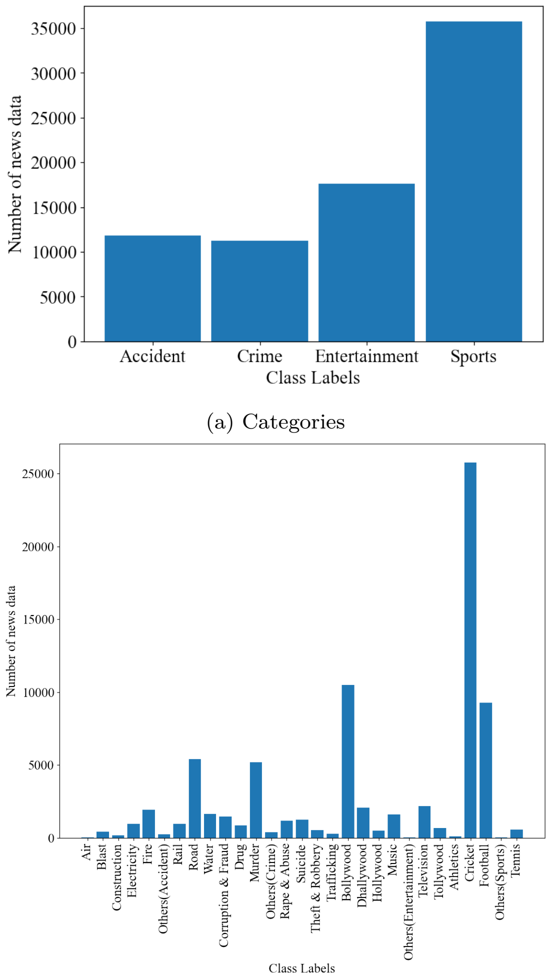
Table 3.6: Data distribution in each news portal

3.3 Proposed Framework of News Article Classification

3.3.1 Embedding Model Preparation

Explain each embedding technique with hyperparameters. Not put general explanation. How to use these technique for your task with various hyperparameters optimisation. Summarize all methods hyperpameters in a table. Done

Word Embeddings are a type of feature extraction technique for selecting a set of features relevant to the input data. It helps in reducing the amount of input data required in training the model by selecting a subset of relatable features. For our research purpose, we have used Keras Embedding layer, Word2vec, and



(a) Categories

(b) Sub-categories

Figure 3.1: Distribution of dataset into categories and sub-categories

FastText for embedding our preprocessed train set before feeding it to the MLP classifier model. We have used TF-IDF feature extraction model with the ML models which estimates how relevant a word is to a document in a collection of documents for ML classsifiers. The embedding models and TF-IDF feature extractor along with the parameters we have used to generate the optimal model are discussed below:

- **Word2Vec**: Word2vec processes text by vectorizing words. Here the tokenized train set was fed as input to the generated Word2Vec model which generated a set of feature matrices. We have used Gensim for extracting features using the Word2Vec model. We have used both Continuous Bags of Words (CBOW) and Skip-gram Word2vec models.

- **FastText**: FastText embeddings exploit subword information and representations are learned of character n-grams, and words represented as the sum of the n-gram vectors [4]. We have applied both Continuous Bag of Words (CBOW) and Skip-gram fastText models using Gensim.
- **Keras Embedding**: The Embedding layer transforms word representations into identical word embedding vectors. To embed text documents using Embedding layers they require to be preprocessed and encoded. So we have fed the preprocessed and encoded train data to the Embedding layer model. Embedding layers come with deep learning libraries PyTorch or Tensorflow. We have used the embedding layer available in Tensorflow Keras with different embedding dimensions.
- **TF-IDF**: Its converts raw documents to a matrix of TF-IDF features. TF-IDF takes tokenized data and converts them to feature matrices to be used by ML models.

Table 3.7 summarizes the parameters used for building the word embedding and feature extraction models along with optimal values.

- **Word2Vec**: Explain with variations: CBOW and Skipgram
- **FastText**: Explain with variations: CBOW and Skipgram
- **Keras Embedding**

3.3.2 Model Preparation

Explain all ML Models and MLP with appropriate parameters. Parameters are summarized in the table Done

For the development of the news classification system we have used MLP, which is a type of Artificial Neural Network. For the purpose of comparing our MLP model we have used several machine learning(ML) models. The preparation of these models are discussed below shortly.

Embedding Technique	Parameter	Description	Optimal Value
Keras Embedding Layer	input_dim	Size of the vocabulary	100,000
	output_dim	Dimension of embedding	800
	input_length	Length of the input sequences	2642
Word2vec & FastText	sg	Training Algorithm:CBOW(0) or Skip-gram(1)	0,1
	size (output_dim)	Dimension of embedding	800
	window	Maximum distance between a target word and words around the target word	5
	min_count	Minimum count of words to consider	5
	input_length	Length of the input sequences	2642
TF-IDF	analyzer	Features (word or character)	'word'
	ngram_range	Boundary of the range of n-values of n-grams	(1, 1)

Table 3.7: Embedding Model Parameters

3.3.3 ML Models

We have used six ML models that are Logistic Regression(LR), Decision Tree(DT), Random Forest(RF), K-Nearest Neighbor(KNN), Naive Bayes(NB) and Support Vector Machine(SVM). Table 3.8 shows the parameters used in developing the ML classifier models used to classify the news documents.

3.3.4 MLP

: Explain here the MLP in details. Summarize hyperparametes in a table. Done

Experimental Model: Move in Chapter 3 Done

We have generated the Multilayer Perceptron(MLP) model with the word embedding models- Keras Embedding layer, Word2vec, and FastText separately for creating the MLP classification framework. Different parameters used in the model are summarized in Table 3.9.

ML Model	Parameter	Value
LR	solver	'lbfgs'
	max_iter	1000
DT	criterion	'gini'
	splitter	'best'
RF	criterion	'gini'
KNN	n_neighbors	2
NB	alpha	1.0
SVM	kernel	'linear'
	random_state	0

Table 3.8: ML Model Summary

Kernel Initializer	HeUniform
No. of hidden layers	1
No. of units(per hidden layer)	450
Dropout rate	0.5
Optimizer	Adam
Learning rate	0.001
Activation function(hidden layer)	ReLU
Batch size	32
Loss Function	Sparse Categorical Crossentropy

Table 3.9: MLP Model Summary

3.3.5 Prediction

The key objective of our work is to develop a news article classification system that can classify the Bangla news articles into appropriate categories and subcategories. To serve our purpose, we have developed a corpus of our own and used both the description and the headline sections for classification. The overall process of the work consists of the following significant steps: train/validation/test split, data preprocessing, word embedding, classifier model generation, model tuning and training, and prediction and evaluation. Figure 3.2 demonstrates the abstract representation of our news classification framework.

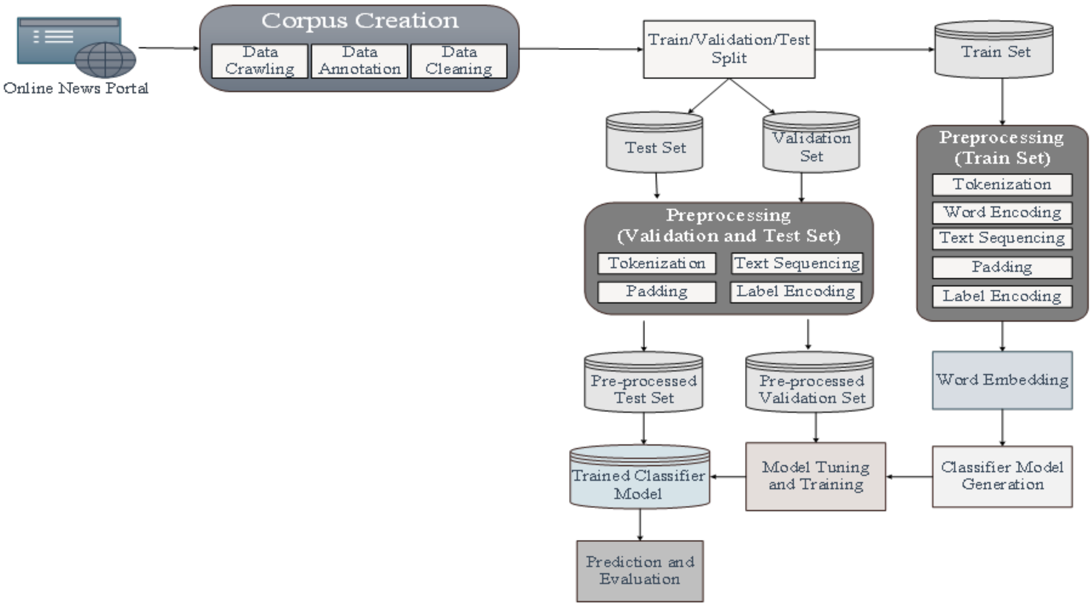


Figure 3.2: Abstract Representation of Proposed News Classification System

3.3.6 News corpus

The news classification model we have developed involves supervised learning which requires a well-defined labeled text dataset. A labeled text corpus consists of a set of text documents along with predefined labels which help to categorize the texts according to the context. These datasets are intended to train models to classify and predict labels of real-world text data. So for the development of an accurate model, a well-established corpus is a fundamental requirement.

There are several datasets available in Bangla language processing to classify news articles but none for sub classification. For the implementation of our news classifier model, we have collected a huge amount of news articles along with their headlines and labeled them into 4 categories. The news documents in each category are further labeled into 29 subcategories. These news articles have been collected from various well-known online news portals. The performance of the classifier model depends mostly on the quality of the corpus.

3.3.7 Train/Validation/Test split

The news text documents (articles and headlines) are split into three different subsets: train, validation, and test, to prevent overfitting. The ratio used here is 80% train, 10% validation, and 10% test.

The train set is the sample of our dataset used to train the classifier model. Since the model learns the weights and biases from the train set it holds the largest proportion of the dataset. The train set we have developed holds 80% of the total text documents. The validation set holds 10% of our text documents, which is the proportion of the dataset used to evaluate the performance of the model while training. It is also used to tune and update the hyperparameters. The validation set affects the model indirectly throughout the development of the model. The test set is the unseen dataset used to evaluate the final model. After the model is trained and tuned using the train and validation set respectively, the test set is used to evaluate the performance of the model based on real-world data. We have used 10% of our dataset as the test set for our model.

Since there are 29 subcategories and the corpus is imbalanced we have used proportional stratified sampling to divide the data in each subcategory proportionally into the train, validation, and test sets.

3.3.8 Data Preprocessing

Data preprocessing defines various processing performed on raw data to prepare it for feeding it to the model. Data preprocessing for Neural Networks includes data cleaning, filling missing values, encoding categorical data into numerical and scaling features into the same range. The preprocessing steps followed are discussed shortly.

Label Encoding: Label Encoding is the process of converting labels into numeral values so that it can be fed to the model. Label Encoding converts text categories into machine-readable form and this is a crucial step for deep learning models. We have encoded the labels by assigning them unique integer values.

Tokenization: Tokenization is the process of slicing a sequence of characters into pieces, called tokens [30]. A token is a string of contiguous characters grouped as a semantic unit and delimited by space, punctuation marks, and newlines. The description and the title sections of the train, validation, and test set were tokenized and converted into tokens.

Word Encoding: Word encoding is the process of transforming words into numbers, which is done my mapping each unique word in the corpus to a particular value, i.e., a number. A pre-defined number of words from the the vocabulary of the tokenized train set has been chosen by limiting the number of most frequent words and then encoded into a word index.

Text Sequencing: Text Sequencing is the process of converting a text document into a list of integers. We have taken the train, validation and test sets and assigned integer values to each word in the document based on the word index generated earlier for the news articles and their headlines separately.

Padding: Generally all the articles are not of the same length but since neural networks require inputs of same size, we have used padding for scaling the lists into same length. We have used post padding where '0' is padded at the end of the sequence to make them of the same length.

Algorithm 2 outlines the overall process of data preprocessing.

3.3.9 Word Embedding

Word Embedding is a type of feature extraction technique for selecting a set of features relevant to the input data. It helps in reducing the amount of input data required in training the model by selecting a subset of relatable features. There are several embedding techniques. For our research purpose, we have used the Keras Embedding layer that comes with the deep learning library Tensorflow. We have used two other feature extraction methods, Word2vec and Fasttext for further tests.

Word2vec is a model that processes text by vectorizing words and create word embeddings. Here the tokenized dataset is given as input into the model and outputs a set of feature matrices. We have used Gensim for extracting features using

Algorithm 2: Data Preprocessing Algorithm

```
Input: TrainSet, ValidationSet, TestSet, LabelSet, MaxLength
Output: TextSequence, LabelSequence
D ←[TrainSet, ValidationSet, TestSet]
L ←[LabelSet]
PDSF ← List of Punctuations, Digits, Stopwords & Foreign words
TextSequence ← list()
for each unique labels $l_i \in LabelSet$ do
    Encode $l_i intoWordIndex$                              ▷ Label Encoding
for each document $d_i \in D$ do
    Remove PDSF                                             ▷ Data cleaning
    TokenizedDocument ← $Tokenizer(d_i)$                    ▷ Tokenization
    WordIndex ← dict()
    for each unique tokens $tr_i \in TokenizedDocument$ and
    $tr_i \in TrainSet$ do
        Encode $tr_i intoWordIndex$                         ▷ Word Encoding
    Sequence ← list()
    for each token $tr_i \in TokenizedDocument$ do
        if $tr_i \in WordIndex$ then
            Sequence.append(WordIndex[$tr_i$])              ▷ Text Sequencing
        else
            Sequence.append(0)
        while length(Sequence) $\neq MaxLength$ do
            Sequence.append(0)                              ▷ Padding
```

the Word2Vec model. We have used both Continuous Bags of Words (CBOW) and Skip-gram Word2vec models. FastText embeddings exploit subword information to construct word embeddings where representations are learned of character n-grams, and words represented as the sum of the n-gram vectors [4]. Here we have applied both Continuous Bag of Words (CBOW) and Skip-gram fastText models using Gensim.

The embedding model maps the vocabulary and reduces it into a feature matrix of dimension (input_length x output_dim) ϵ (2642 x 800) which is then fed to the classifier model.

3.3.10 Classifier Model Generation

In our research, we have used Multilayer Perceptron (MLP) as the classifier model, which is a class of feed-forward artificial neural networks. MLPs consist of an input layer, an output layer, and one or more hidden layers in between these two layers. The neurons in the hidden layers perform the classification of the features with the help of non-linear activation functions and predefined weights and bias. The architecture of the classifier model is shown in Figure 3.3.

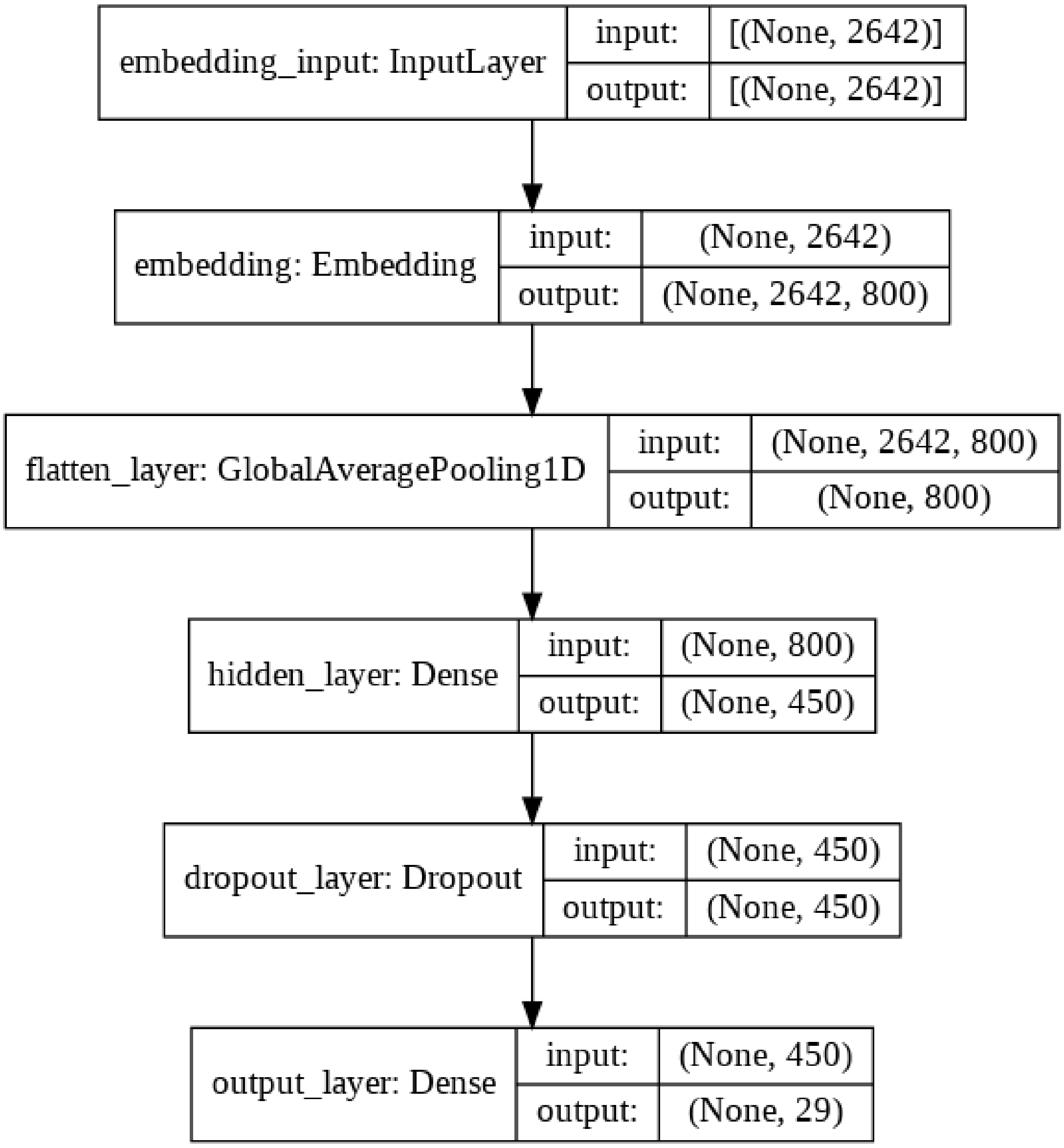


Figure 3.3: Classifier Model Architecture

The InputLayer is the layer that takes in the processed train set as input and passes them to the adjacent layer of the model. Since no processing is done in this layer, the input and output shapes are the same which is (None, input_length) ϵ (None, 2642). Here input_length is the length of the input preprocessed data. The Embedding layer works as the feature extraction model which outputs a feature matrix of shape (None, input_length, output_dim) ϵ (None, 2642, 800) where input_length is the length of the input data and output_dim is the dimension of the embedding. For word2vec and fastText models this layer takes the weights calculated through these embedding models. The matrix is then flattened into a one-dimensional vector of shape (None,output_dim) ϵ (None, 800) using the GlobalAveragePooling1D layer. The flattened vector then passes to the Dense layer which is a hidden layer with 450 units that uses the 'relu' activation function to learn the distinguishing learning parameters. To avoid overfitting dropout rate of 0.5 is used through the Dropout layer. The output shape of this layer is (None, units) ϵ (None, 450). The final Dense layer which is the output layer gives the output prediction results of shape (None, units) ϵ (None, 29). It has units equal to the number of categories and uses the 'softmax' activation function for predicting probabilities of the class labels.

3.3.11 Model Tuning and Training

The generated model is first compiled using the 'Adam' optimizer with a learning rate of '0.001' and the 'sparse_categorical_crossentropy' loss function. For tuning the model we have used 'accuracy' as the metric. The preprocessed train set and validation set along with their encoded labels are fed to the compiled model for training and tuning. We have adopted a set of necessary hyperparameters for tuning the model to get the optimal value of the parameters. The hyperparameter optimization and the training process are described below shortly.

Hyperparameter Optimization: Hyperparameter optimization is an important aspect in terms of deep learning models. Since are a lot of hyperparameters on which the performance of the models depends, the models are needed to be tuned carefully to get the optimal hyperparameter. The choice of hyperparameters can make the difference between poor and superior predictive accuracy. There are several ways for tuning hyperparameters. Grid search can be applied for tuning all possible combinations of hyperparameters at once when there are enough computational resources available. But it has been observed that models can be tuned in less time through randomly chosen values of hyperparameters rather than exhaustive grid search [31]. The MLP model along with the feature extraction model has been tuned with different hyperparameters, the result of which is summarized in Table 3.10.

Model Training: After the model is tuned through hyperparameter optimization, the optimal values are used for training the model. The classifier model is trained by applying the optimal set of parameters and using the train set to create the trained classifier model.

Hyper-parameter	**Search Space**	**Optimal Value**
Learning Rate	[0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001]	0.001
Optimizer	[Adam, Adamax, Adagrad, Adadelta, Nadam, RMSprop, SGD, Ftrl]	Adam
Activation function	[ReLU, tanh, sigmoid, softmax, softplus, softsign, selu, elu]	ReLU
Dropout	[0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	0.5
No. of Hidden Layers	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	1
No. of Units (per hidden layer)	[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 350, 500, 550, 600, 650, 700, 750, 800, 900, 1000]	900
Batch Size	[1, 32, 64, 128, 256, 1024, 61122]	32
Embedding Dimension	[10, 50, 100, 150, 200, 250, 300, 350, 400, 500, 600, 700, 800]	800
Vocab Size	[5k, 10k, 30k, 50k, 80k, 100k ,330k]	10k

Table 3.10: Hyper-parameter tuning

3.3.12 Prediction and Evaluation

The trained classifier model is used for predicting labels of the unseen test set data. Class labels of the unlabeled test set are used for the evaluation of the trained classifier model. 7592 news articles of the pre-processed test set are fed to the trained model, which then predicts the label of the news articles using the ’softmax’ probability distribution. The following equation is used to calculate the probability distribution-

$$Softmax(\theta_i) = \frac{exp(\theta_i)}{\sum_{i=1}^{n} exp(\theta_i)} \tag{3.1}$$

Here θ_i is the output feature vector from the trained model, and n is the number of categories. The output values range from 0 to 1 and the class one with the highest probability is taken as the predicted label. Comparing the predicted output labels with the actual labels, the trained classifier model is evaluated.

Add another Chapter 4: Implementation Details Done

3.4 Introduction

3.5 System Requirements

Combine h/w and s/w here.

3.6 Keras Framework

Explain how to use/prepare Keras Framework

3.7 System Set up and Runing

Explain how to set up and run the model with test data.

3.7.1 Implementation Snapshot

Use few sample input/out with system screen shot.

3.8 Impact Analysis

Follow Template

3.9 Conclusion

Move this part in the Implementation Chapter 4 Done

3.10 Conclusion

Throughout this chapter, we have explained the overall system of news classification. Each and every module of our proposed methodology is described briefly. The overall sytem architecture is given in Figure 3.2. We have also mentioned an algorithm used for data preprocessing. The generated classifier model is described through a graph in Figure 3.3.

Chapter 4

Implementation

4.1 Introduction

In this chapter the implementation of our system is explained. The hardware and software requirements used for implementation of our system are also mentioned in this chapter. The impact of our proposed system is also mentioned in this chapter.

4.2 System Requirements

For the implementation of our system a few hardware and software tools are required. The required hardware and software are listed below.

4.2.1 Hardware Requirements

The experiments have been performed on a general-purpose computer with the following specifications:

- Intel® Core™ i3-5005U CPU @ 2.00GHz
- RAM 4.0GB

For the development of our system, we have used Google Colab. The hardware specification provided in the Google Colab environment are:

- Cloud TPU v2 64GB
- Intel(R) Xeon(R) CPU @ 2.30GHz
- RAM 12.6 GB
- Disk 33 GB

4.2.2 Software Requirements

We have implemented our system in a specific software environment using the following software tools, languages and libraries:

- Operating System 64-bit Windows 10 Pro
- Jupyter Notebook (anaconda 3)
- Notebook++
- Python version 3.6
- TensorFlow version 2.4.1
- Keras
- Numpy
- Pandas
- Matplotlib

4.3 Keras Framework

Deep learning methods can be maintained using different libraries such as Theano, TensorFlow, Caffe, Mxnet, etc. Keras is one of the most robust and easy-to-maintain APIs which is built on top of popular libraries like TensorFlow for constructing deep learning models. TensorFlow is adaptable and it promotes distributed computing. To use Keras as a backend we have followed the steps given below:

- Create a virtual environment using virtualenv.
- Activate the environment.
- Install the libraries required to run the models, like NumPy, Pandas, etc.

4.4 System Set up and Running

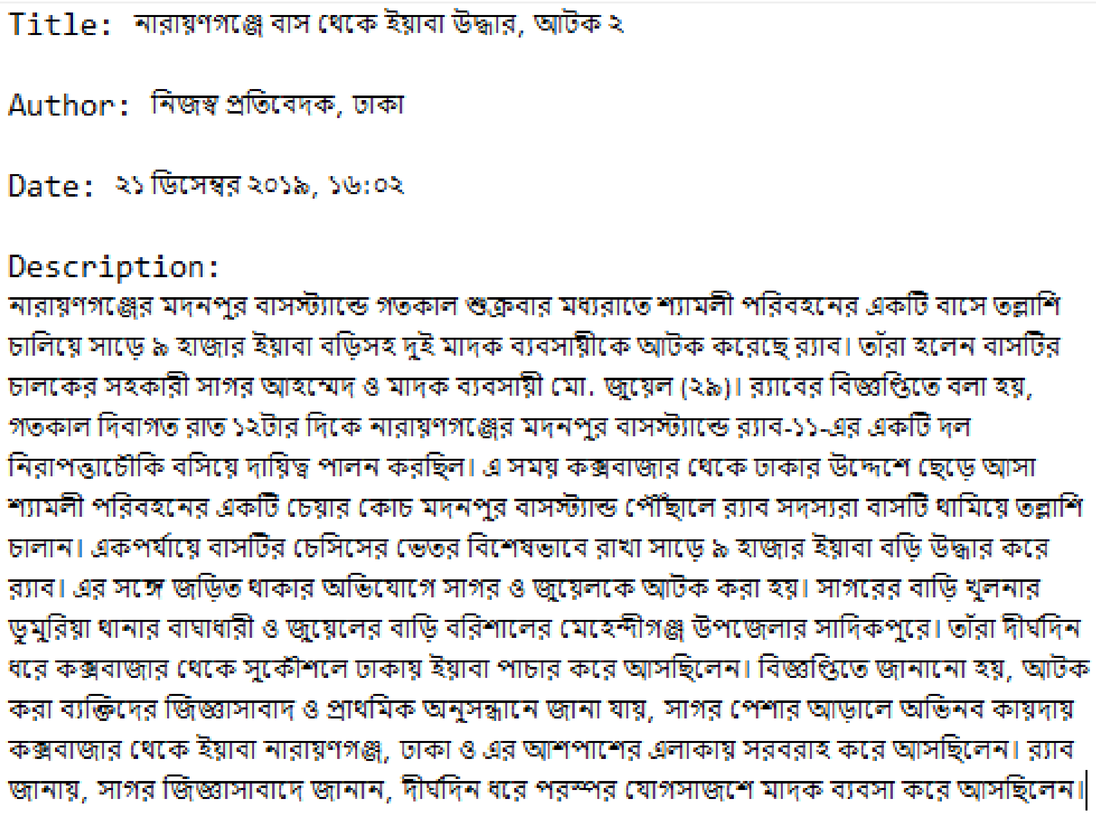
Explain how to set up and run the model with test data. Done.

The whole system was set up and run on Google Colab. Keras with TensorFlow was installed as the backend. The training, tuning and testing of the overall model was completed utilizing TPU available within Google Colab. The trained model was saved in local disk which was then used to further test the trained model using test data.

4.4.1 Implementation Snapshot

Use few sample input/out with system screen shot. Done.

Figure 4.1 gives a view of our original news dataset which consists of Title, Author, Date and Description. For implementation of our classifier model we have used both the Title and the Deription section where Title depicts the headline and Description depicts the article of the news.



Title: নারায়ণগঞ্জে বাস থেকে ইয়াবা উদ্ধার, আটক ২

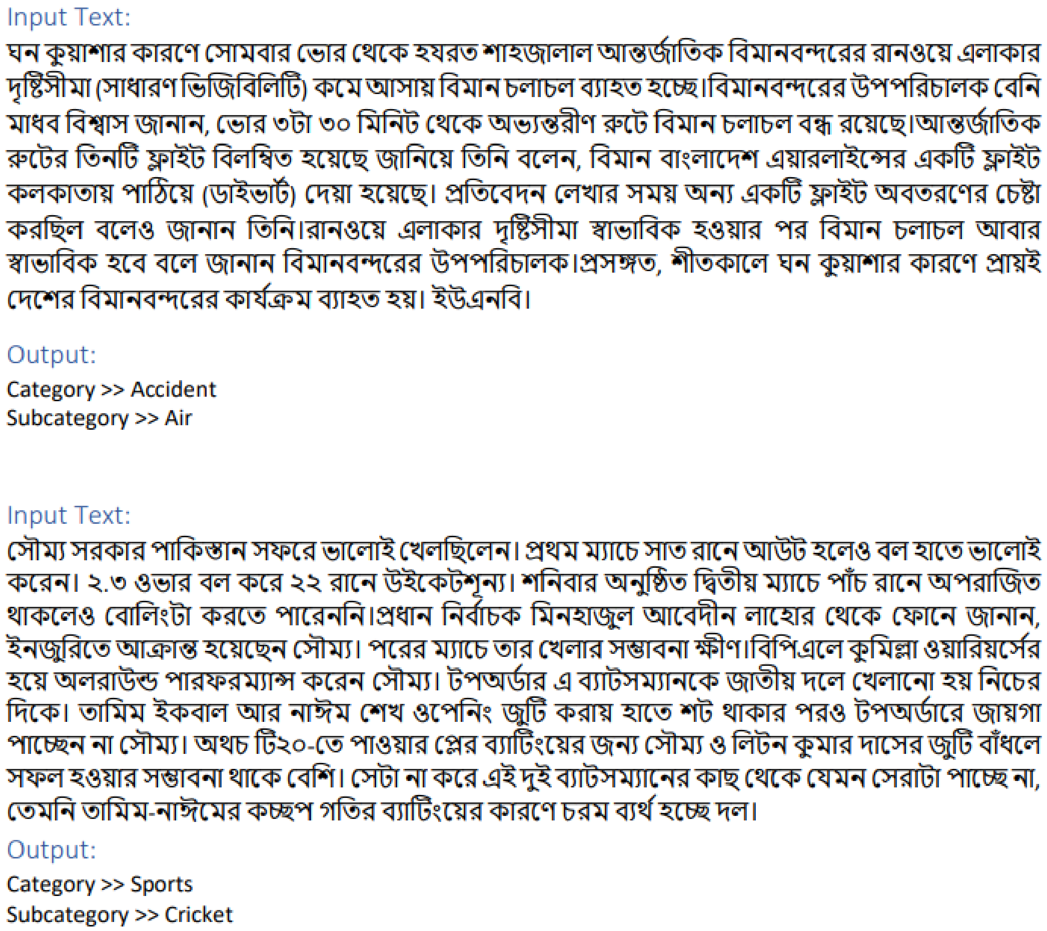
Author: নিজস্ব প্রতিবেদক, ঢাকা

Date: ২১ ডিসেম্বর ২০১৯, ১৬:০২

Description:
নারায়ণগঞ্জের মদনপুর বাসস্ট্যান্ডে গতকাল শুক্রবার মধ্যরাতে শ্যামলী পরিবহনের একটি বাসে তল্লাশি চালিয়ে সাড়ে ৯ হাজার ইয়াবা বড়িসহ দুই মাদক ব্যবসায়ীকে আটক করেছে র‍্যাব। তাঁরা হলেন বাসটির চালকের সহকারী সাগর আহম্মেদ ও মাদক ব্যবসায়ী মো. জুয়েল (২৯)। র‍্যাবের বিজ্ঞপ্তিতে বলা হয়, গতকাল দিবাগত রাত ১২টার দিকে নারায়ণগঞ্জের মদনপুর বাসস্ট্যান্ডে র‍্যাব-১১-এর একটি দল নিরাপত্তাচৌকি বসিয়ে দায়িত্ব পালন করছিল। এ সময় কক্সবাজার থেকে ঢাকার উদ্দেশে ছেড়ে আসা শ্যামলী পরিবহনের একটি চেয়ার কোচ মদনপুর বাসস্ট্যান্ড পৌঁছালে র‍্যাব সদস্যরা বাসটি থামিয়ে তল্লাশি চালান। একপর্যায়ে বাসটির চেসিসের ভেতর বিশেষভাবে রাখা সাড়ে ৯ হাজার ইয়াবা বড়ি উদ্ধার করে র‍্যাব। এর সঙ্গে জড়িত থাকার অভিযোগে সাগর ও জুয়েলকে আটক করা হয়। সাগরের বাড়ি খুলনার ডুমুরিয়া থানার বাঘাধারী ও জুয়েলের বাড়ি বরিশালের মেহেন্দীগঞ্জ উপজেলার সাদিকপুরে। তাঁরা দীর্ঘদিন ধরে কক্সবাজার থেকে সুকৌশলে ঢাকায় ইয়াবা পাচার করে আসছিলেন। বিজ্ঞপ্তিতে জানানো হয়, আটক করা ব্যক্তিদের জিজ্ঞাসাবাদ ও প্রাথমিক অনুসন্ধানে জানা যায়, সাগর পেশার আড়ালে অভিনব কায়দায় কক্সবাজার থেকে ইয়াবা নারায়ণগঞ্জ, ঢাকা ও এর আশপাশের এলাকায় সরবরাহ করে আসছিলেন। র‍্যাব জানায়, সাগর জিজ্ঞাসাবাদে জানান, দীর্ঘদিন ধরে পরস্পর যোগসাজশে মাদক ব্যবসা করে আসছিলেন।

Figure 4.1: Random sample data

The prediction of our model using news articles and news headlines are shown in Figure 4.2 and Figure 3.3 respectively.



Input Text:

ঘন কুয়াশার কারণে সোমবার ভোর থেকে হযরত শাহজালাল আন্তর্জাতিক বিমানবন্দরের রানওয়ে এলাকার দৃষ্টিসীমা (সাধারণ ভিজিবিলিটি) কমে আসায় বিমান চলাচল ব্যাহত হচ্ছে।বিমানবন্দরের উপপরিচালক বেনি মাধব বিশ্বাস জানান, ভোর ৩টা ৩০ মিনিট থেকে অভ্যন্তরীণ রুটে বিমান চলাচল বন্ধ রয়েছে।আন্তর্জাতিক রুটের তিনটি ফ্লাইট বিলম্বিত হয়েছে জানিয়ে তিনি বলেন, বিমান বাংলাদেশ এয়ারলাইন্সের একটি ফ্লাইট কলকাতায় পাঠিয়ে (ডাইভার্ট) দেয়া হয়েছে। প্রতিবেদন লেখার সময় অন্য একটি ফ্লাইট অবতরণের চেষ্টা করছিল বলেও জানান তিনি।রানওয়ে এলাকার দৃষ্টিসীমা স্বাভাবিক হওয়ার পর বিমান চলাচল আবার স্বাভাবিক হবে বলে জানান বিমানবন্দরের উপপরিচালক।প্রসঙ্গত, শীতকালে ঘন কুয়াশার কারণে প্রায়ই দেশের বিমানবন্দরের কার্যক্রম ব্যাহত হয়। ইউএনবি।

Output:

Category >> Accident
Subcategory >> Air

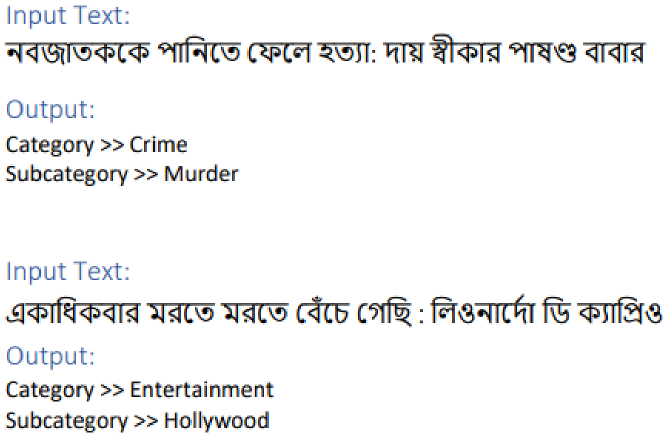
Input Text:

সৌম্য সরকার পাকিস্তান সফরে ভালোই খেলছিলেন। প্রথম ম্যাচে সাত রানে আউট হলেও বল হাতে ভালোই করেন। ২.৩ ওভার বল করে ২২ রানে উইকেটশূন্য। শনিবার অনুষ্ঠিত দ্বিতীয় ম্যাচে পাঁচ রানে অপরাজিত থাকলেও বোলিংটা করতে পারেননি।প্রধান নির্বাচক মিনহাজুল আবেদীন লাহোর থেকে ফোনে জানান, ইনজুরিতে আক্রান্ত হয়েছেন সৌম্য। পরের ম্যাচে তার খেলার সম্ভাবনা ক্ষীণ।বিপিএলে কুমিল্লা ওয়ারিয়র্সের হয়ে অলরাউন্ড পারফরম্যান্স করেন সৌম্য। টপঅর্ডার এ ব্যাটসম্যানকে জাতীয় দলে খেলানো হয় নিচের দিকে। তামিম ইকবাল আর নাঈম শেখ ওপেনিং জুটি করায় হাতে শট থাকার পরও টপঅর্ডারে জায়গা পাচ্ছেন না সৌম্য। অথচ টি২০-তে পাওয়ার প্লের ব্যাটিংয়ের জন্য সৌম্য ও লিটন কুমার দাসের জুটি বাঁধলে সফল হওয়ার সম্ভাবনা থাকে বেশি। সেটা না করে এই দুই ব্যাটসম্যানের কাছ থেকে যেমন সেরাটা পাচ্ছে না, তেমনি তামিম-নাঈমের কচ্ছপ গতির ব্যাটিংয়ের কারণে চরম ব্যর্থ হচ্ছে দল।

Output:

Category >> Sports
Subcategory >> Cricket

Figure 4.2: System Implementation (news articles)



Input Text:

নবজাতককে পানিতে ফেলে হত্যা: দায় স্বীকার পাষণ্ড বাবার

Output:

Category >> Crime
Subcategory >> Murder

Input Text:

একাধিকবার মরতে মরতে বেঁচে গেছি : লিওনার্দো ডি ক্যাপ্রিও

Output:

Category >> Entertainment
Subcategory >> Hollywood

Figure 4.3: System Implementation (news headlines)

4.5 Impact Analysis

Text classification is one of the essential elements of text analysis. News classification is a field of text classification that labels natural language texts with relevant predefined classes. News classification systems have several impacts which are described below shortly.

4.5.1 Social and Environmental Impact

Due to the messy nature of the text analyzing, understanding, organizing, and sorting through text data is hard and time-consuming so most organizations fail to extract value from that. So, classifying them into different categories by sorting, managing and filtering can help in organizing text documents. The same is the case with news articles. An automated news classification system can help different organizations that require information from news articles to find the necessary information. It can help online news portals in organizing their daily published news articles without much human maintenance.

4.5.2 Ethical Impact

News classification systems can help law practitioners, social researchers, government, and other organizations to organize and manage news related to crime and other activities. Our system can help researchers to find different crime rates based on the crime news.

4.6 Conclusion

The implementation tools required throughout the system are mentioned in this chapter along with the impact it has. The Keras Framework that we have used is also explained in this chapter.

Chapter 5

Results and Discussions

5.1 Introduction

Chapter 5 explains the details experimental analysis on developed corpus for various ML methods. This Chapter also investigates the performance of implemented models with various evaluation measures (such as precision, recall, accuracy, $f1$-score). The details comparative analysis among various methods with existing techniques also illustrated in this Chapter. For better insight, a details error analysis of the proposed model is included at end of this Chapter.

Mode all relevant parts to Chapter 3. Done

5.2 Experiments

Add few sentences about the purpose of the experiment. Done The experiments are done on the test data to give an overview of the performance of our news classification model. The experiment evaluates the MLP news classifier model along with other ML models to give the illustration of how good the model performs on unseen real world news data.

5.2.1 Experimental Data

Presets here the data split up in a table: train/test and validation set with proper explanation. Shows the each set distribution with class-wise and sub-class wise. Done

For the purpose of experimentation the whole data has been split up into train/validation/test set. The train set holds 80% of the whole dataset since the model training depends on it. The model is trained with the train set consisting of 61122 news articles. 10% of the dataset is used for tuning the model known as the validation set. To validate the model at each epoch while training, we have used the validation set of 7629 news articles. The other 10% is the test set that is used for evaluating the trained classifier model. The test set with 7529 news articles has been used for the evaluation of the trained model. The data split up is shown in Table **??** for each categories and subcategories.

The summary of the experimental news article and headline data is shown in Table 5.2.

Section	**Criteria**	**Train Set**	**Validation Set**	**Test Set**	**Total**
News Article	Number of articles	61,122	7,629	7,592	76,343
	Number of words	9,485,822	1,173,586	1,184,003	11,843,411
	Unique Words	331,529	95,769	96,105	523,403
Headline	Number of articles	61,122	7,629	7,629	76,343
	Number of words	299,291	37,344	37,225	373,860
	Unique Words	38,108	11,722	11,640	61,470

Table 5.2: Summary of experimental data

5.2.2 Evaluation Measures

The overall classification model was evaluated in two different phases: train/validation phase and test phase. Several measures such as confusion matrix (CM), accuracy (A), loss, precision (P), recall (R0, and F_1-score are considered for model's evaluation.

- **Confusion matrix:** The confusion matrix is a evaluation tool used for

Category	Train	Validation	Test	Subcategory	Train	Validation	Test
Accident	9487	1183	1171	Air	42	5	4
				Blast	343	43	41
				Construction	154	19	18
				Electricity	769	96	94
				Fire	1540	192	191
				Rail	782	97	96
				Road	4342	542	541
				Water	1320	165	163
				Others	195	24	23
Crime	8992	1120	1110	Corruption & Fraud	1173	146	145
				Drug	682	85	83
				Murder	4166	520	159
				Rape & Abuse	950	118	177
				Suicide	1018	127	128
				Theft & Robbery	455	55	54
				Trafficking	230	28	27
				Others	328	41	39
Entertainment	17568	14065	1756	Bollywood	8398	1050	1048
				Dhallywood	1654	206	205
				Hollywood	396	49	48
				Music	1285	160	159
				Television	1738	217	216
				Tollywood	556	69	68
				Others	38	5	3
Sports	28578	3570	3564	Athletics	75	9	8
				Cricket	20616	2577	2575
				Football	7413	926	925
				Tennis	454	56	55
				Others	20	2	1

Table 5.1: Data Split Up into train/validation/test set

error analysis of the trained classifier model using the test set. It is a performance measuring matrix used for classification models where the actual

		Total Words		**Unique Words**	
Class	**Data**	**Article**	**Headline**	**Article**	**Headline**
Accident	11841	2,077,659	74,387	100,018	6,834
Crime	11,222	2,998,583	78,160	127,494	10,175
Entertainment	17,568	2,828,029	120,263	166,258	20174
Sports	35,712	7,751,245	243,971	269,226	22,984
Total	**76,343**	**15,655,516**	**516,781**	**662,996**	**60,167**

Table 5.3: Data statistics in each class

labels of the test set are known. It evaluates the performance by making predictions on the test set. It is also known as an error matrix since it reveals the errors that occurred by the model while predicting labels of the test set. The matrix is divided into two dimensions, the true labels, and the predicted labels. It evaluates the performance of the classification model There are four cases related to confusion matrices. These are:

- True Positive (TP): correct positive prediction.
- False Positive (FP): incorrect positive prediction.
- True Negative (TN): correct negative prediction.
- False Negative (FN): incorrect negative prediction.

- **Accuracy:** Accuracy (ACC) estimates the portion of accurate predictions. It ranges from zero to one. A larger value of accuracy means better prediction made by the model. The best accuracy calculated is 1.0 and the worst is 0.0. The accuracy(ACC) is calculated using Eq-5.1.

$$ACC \ = \ \frac{TP \ + \ TN}{TP + FP + TN + FN} \tag{5.1}$$

where TP, FP, TN and FN are true positive, false positive, true negative, and false negative, respectively.

- **Loss:** We used the 'sparse_categorical_crossentropy' loss function to calculate the Loss(L) by computing the following sum:

$$L \;=\; -\frac{1}{n}\sum_{j=1}^{n} y^{(j)}\; log\; (\bar{y}^{(j)}) \tag{5.2}$$

where $y^(j)$ is the actual class label and $\bar{y}^(j)$ is the predicted class label of the jth article.

- **Precision:** Precision computes the portion of true positives compared to the total predicted positives. It can range from zero to one. A larger value of precision means better predictive accuracy. The best precision value is 1.0 and the worst is 0.0. The precision(P) is calculated using Eq-5.3.

$$P \;=\; \frac{TP}{TP + FP} \tag{5.3}$$

where TP and FP are true positive and false positive, respectively.

- **Recall:** Recall also known as sensitivity, measures the fraction of true positives compared to the total predicted positives. It can range from zero to one. A larger value of recall indicates better performance. The best recall value is 1.0 and the worst is 0.0. The recall(R) is calculated using Eq-5.4.

$$R \;=\; \frac{TP}{TP + FN} \tag{5.4}$$

where TP and FN are true positive and false negative, respectively.

- *F*__1**-score:** F1-score helps in evaluating recall and precision at the same time when one of them is high and the other is low. The higher value of the F1 score indicates that recall and precision are almost equal. The F1-score(F1) is calculated using Eq-5.5.

$$F1 \;=\; 2\frac{P * R}{P + R} \tag{5.5}$$

where P and R are precision and recall, respectively.

5.2.3 Results Analysis

First we have performed an performance analysis on 5 different word embedding models that are Keras Embedding Layer, Word2vec CBOW, Word2vec Skip-gram, FastText CBOW and FastText Skip-gram which are summarized in Table 5.4. Here we have evaluated the word embedding techniques based on accuracy using the parameters given in Table 3.7 for the embedding models and Table 3.9 for the MLP model.

I am not understand why you ignored other measures, include them so. Sir I have used other measures such as precision, recall, f1 score for Keras embedding layer that works the best. Sir is that okay? or should I add these measures in Table 5.4 also?

Section	**Word Embedding Model**	**Accuracy(%) for 4 categories**	**Accuracy(%) for 29 subcategories**
News Article	Keras Embedding Layer	**98.18**	**90.63**
	Word2vec(CBOW)	97.47	86.97
	Word2vec(Skip-gram)	97.63	88.96
	FastText(CBOW)	97.21	86.35
	FastText(Skip-gram)	97.21	86.62
Headline	Keras Embedding Layer	**94.53**	**82.97**
	Word2vec(CBOW)	80.81	51.69
	Word2vec(Skip-gram)	92.19	58.40
	FastText(CBOW)	80.93	49.01
	FastText(Skip-gram)	80.59	52.09

Table 5.4: Classification Results

From Table 5.4 we can observe that our MLP model achieved the highest accuracy of 98.18% (for 4 categories) and 90.32% (for 29 sub-categories) using Keras

Embedding Layer for news articles classification. When the headlines are used for classification highest accuracy of 94.53% (for 4 categories) and 86.35% (for 29 sub-categories) using Keras Embedding Layer. Overall, we can say that the classification model achieves the highest accuracy when Keras Embedding Layer is used as the feature extraction model for both category and subcategory classification.

Performance of MLP news classification model with Keras Embedding layer for category and subcategory classification using news article and headline classification is shown in Table 5.5 and Table 5.6 respectively.

Category Name	**News Article**			**Headline**			**No. of test data**
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	
Sports	0.99	0.98	0.99	0.97	0.96	0.97	3564
Entertainment	0.95	0.99	0.97	0.90	0.94	0.92	1747
Accident	0.97	0.99	0.98	0.96	0.92	0.94	1171
Crime	0.99	0.97	0.98	0.91	0.92	0.92	1110
Weighted Average	0.98	0.98	0.98	0.95	0.95	0.95	7592

Table 5.5: Model performance of test data(4 categories)

From Table 5.5 we can observe that for both news article and headline classification Sports category has the highest F1-score. Table 5.6 shows that for both news article and headline classification subcategory Cricket under the Sports category has the highest F1-score. It can be seen that the Others subcategory under Entertainment and the Others subcategory under Sports has a 0.0 F1-score and these two subcategories have the least amount of data. From both these tables, it can be said that classification using news articles gives more accuracy compared to the headlines since articles have more words compared to the headline section

Class Name	News Article			Headline			No. of test data
	Precision	Recall	F1-score	Precision	Recall	F1-score	
Cricket	0.99	0.96	0.97	0.92	0.95	0.94	2575
Bollywood	0.84	0.92	0.88	0.75	0.86	0.80	1048
Football	0.95	0.96	0.95	0.89	0.88	0.88	925
Road	0.95	0.96	0.96	0.86	0.93	0.89	541
Murder	0.86	0.92	0.89	0.77	0.79	0.78	519
Television	0.81	0.66	0.73	0.55	0.40	0.46	216
Dhallywood	0.72	0.82	0.77	0.63	0.59	0.61	205
Fire	0.94	0.92	0.93	0.88	0.84	0.86	191
Water	0.93	0.96	0.95	0.89	0.81	0.85	163
Music	0.78	0.75	0.77	0.60	0.58	0.59	159
Corruption & Fraud	0.74	0.81	0.77	0.67	0.70	0.68	145
Suicide	0.91	0.81	0.86	0.66	0.60	0.63	126
Rape & Abuse	0.82	0.82	0.82	0.85	0.66	0.74	117
Rail	0.91	0.93	0.92	0.92	0.75	0.83	96
Electricity	0.94	0.95	0.94	0.77	0.93	0.84	94
Drug	0.84	0.84	0.84	0.73	0.75	0.74	83
Tollywood	0.58	0.56	0.57	0.48	0.24	0.32	68
Tennis	0.88	0.76	0.82	0.86	0.55	0.67	55
Theft & Robbery	0.68	0.48	0.57	0.66	0.57	0.61	54
Hollywood	0.51	0.50	0.51	0.52	0.29	0.37	48
Blast	0.91	0.76	0.83	0.74	0.56	0.64	41
Others(Crime)	0.48	0.36	0.41	0.29	0.13	0.18	39
Trafficking(Accident)	0.75	0.67	0.71	0.83	0.70	0.76	27
Others	0.45	0.43	0.44	0.64	0.30	0.41	23
Construction	0.73	0.44	0.55	0.53	0.56	0.54	18
Athletics	0.83	0.62	0.71	1.00	0.12	0.22	8
Air	0.75	0.75	0.75	0.50	0.25	0.33	4
Others (Entertainment)	0.00	0.00	0.00	0.00	0.00	0.00	3
Others(Sports)	0.00	0.00	0.00	0.00	0.00	0.00	1
Weighted Average	0.91	0.90	0.90	0.82	0.83	0.82	7592

Table 5.6: Model performance of test data(29 subcategories)

which helps the classifier model with more features to distinguish classes from one another. As the number of categories decreases the performance of the classification model increases. It can also be observed that as the number of training data decreases the F1-score of the subcategories also decreases gradually.

5.2.4 Comparison with ML Baselines

What feature extraction or embedding is used for ML? Explain this with reason. It is better to evaluate all combinations of ML + feature extraction/embedding and choose best one in each ML to compare with MLP.

To evaluate the effectiveness of our model, we have compared our proposed system with different machine learning classifier models using TF-IDF feature extraction model with the parameters given in Table 3.7. The classifiers we have compared our system with are Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Naive Bayes, and Support Vector Machine using the parameters mentioned in Table 3.8 for each ML models. Table 5.7 and Table 5.8 summarizes the performance of the news classifiers for categories (4) and subcategories (29). The parameters used in the MLP model is mentioned in Table 3.9. It can be observed that the MLP classifier model using Keras embedding layer outperforms the machine learning approaches using TF-IDF in terms of accuracy. So, it can be said that our proposed news classifier model outperforms ML models for both category classification and subcategory classification whether using news articles or headlines.

Method	**Accuracy(%)**	
	News Article	**News Headline**
Logistic Regression	97.92	92.77
Decision Tree	93.77	88.31
Random Forest	97.22	90.61
K-Nearest Neighbor	54.08	59.68
Naive Bayes	97.07	92.88
Support Vector Machine	98.01	93.82
Multilayer Perceptron[**Proposed**]	**98.18**	**94.53**

Table 5.7: Performace comparison with other models for category(4) classification

Method	**Accuracy(%)**	
	News Article	**News Headline**
Logistic Regression	89.20	78.64
Decision Tree	84.07	73.11
Random Forest	83.86	78.29
K-Nearest Neighbor	41.81	46.31
Naive Bayes	70.18	70.00
Support Vector Classifier	90.54	80.95
Multilayer Perceptron[**Proposed**]	**90.63**	**82.97**

Table 5.8: Performace comparison with other models for subcategory(29) classification

5.3 Error Analysis

To analyze the performance and understand the effectiveness of the classifier model confusion matrix is introduced, where actual labels are compared with the predicted labels concerning the test set.

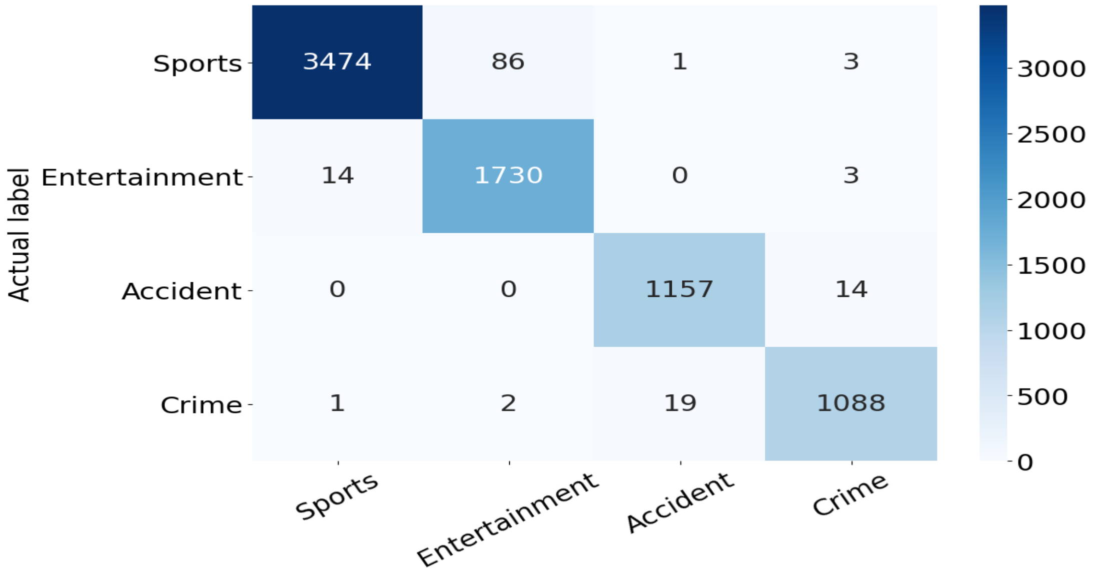


Figure 5.1: Confusion Matrix using news articles(4 categories)

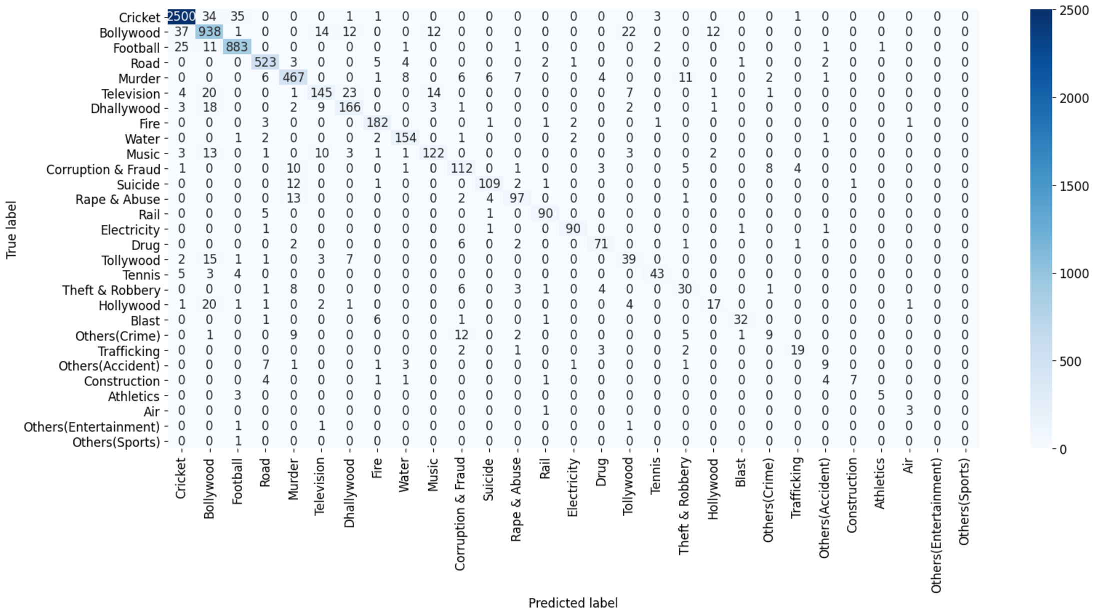


Figure 5.2: Confusion Matrix using news articles(29 subcategories)

Figure 5.1 depicts the confusion matrix regarding the news article classifier(4 categories) for the test set. Here the category Sports gets the most accurate predictions with 3467 True Positives out of 3564 data. But 86 Sports data are being missclassified as Entertainment data, which is the majority of miss-prediction. Since some Sports news such as "বাংলাদেশ জাতীয় ক্রিকেট দলের বাঁ-হাতি ওপেনার সৌম্য সরকার খুলনার মেয়ে প্রিয়ন্তি দেবনাথ পূজার সঙ্গে গাঁটছড়া বেঁধেছেন।" (Bangladesh national cricket team left-handed opener Soumya Sarkar tied the knot with Khulna's daughter Priyanti Debnath Pooja.), includes Entertainment contents like wedding, which often can get mixed up by the classifier. Again Crime articles like "ট্রেনের ছাদ থেকে যুবকের লাশ উদ্ধার। পুলিশ ধারণা করছে, ট্রেনের ছাদে চলন্ত অবস্থায় গাছ বা ওভারব্রিজ কোনো কিছুর সঙ্গে আঘাত পেয়ে ওই যুবকের মৃত্যু হয়েছে।" (The body of the youth was recovered from the roof of the train. Police believe the youth died after being hit by a tree or an overbridge while moving on the roof of the train.) which can also be classified as Accident leads to 19 misclassified Crime news.

Figure 5.2 represents the confusion matrix of the news article classifier(29 categories), where it can be observed that 37 out of 1048 Bollywood(Entertainment) articles are labeled as Cricket(Sports) because of news like "সামনে নিউজিল্যান্ডের বিপক্ষে ভারতের ২ ম্যাচ টেস্ট সিরিজ। এর আগেই বিরাটের সঙ্গে সময় কাটিয়ে দেশে ফিরে আসছেন বলিউড অভিনেত্রী আনুশকা।" (India's 2-match Test series against New Zealand ahead. Bollywood actress Anushka is already returning home after spending time with Virat.). Others(Entertainment) has 0 true predictions out of 3 test data, 2 of which are predicted as Television(Entertainment) and Tollywood(Entertainment) and 1 as Football(Sports). Others(Sports) have 0 true predictions out of 1 test data which is misclassified as Football(Sports). Since the Others(Entertainment) and Others(Sports) subcategories had only 46 and 23 data each, the classifier was not trained properly for these 2 subcategories as deep learning classifiers require more training data.

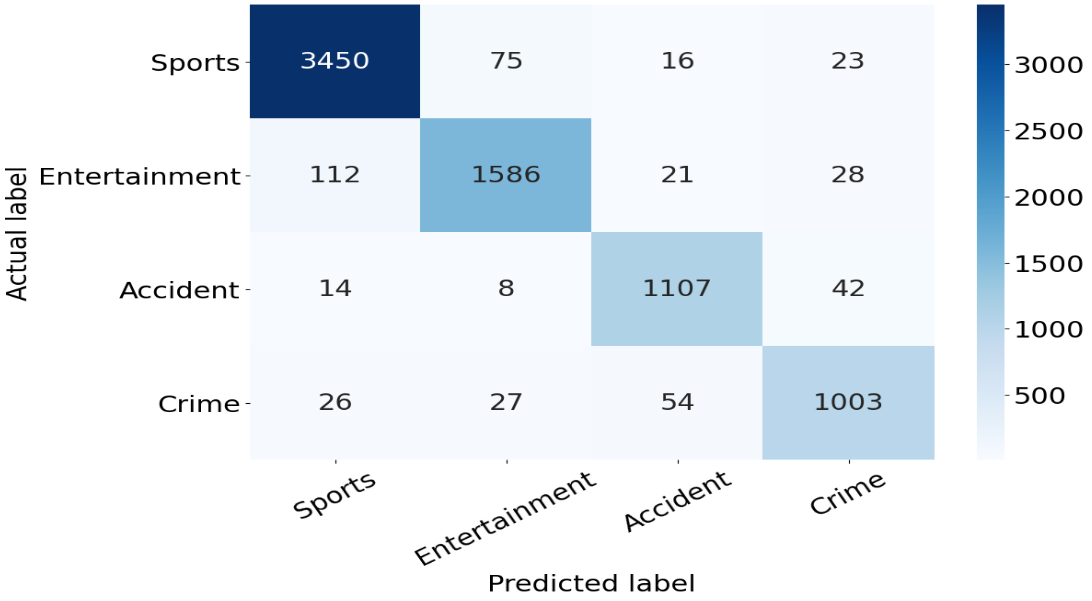


Figure 5.3: Confusion Matrix using news headlines(4 categories)

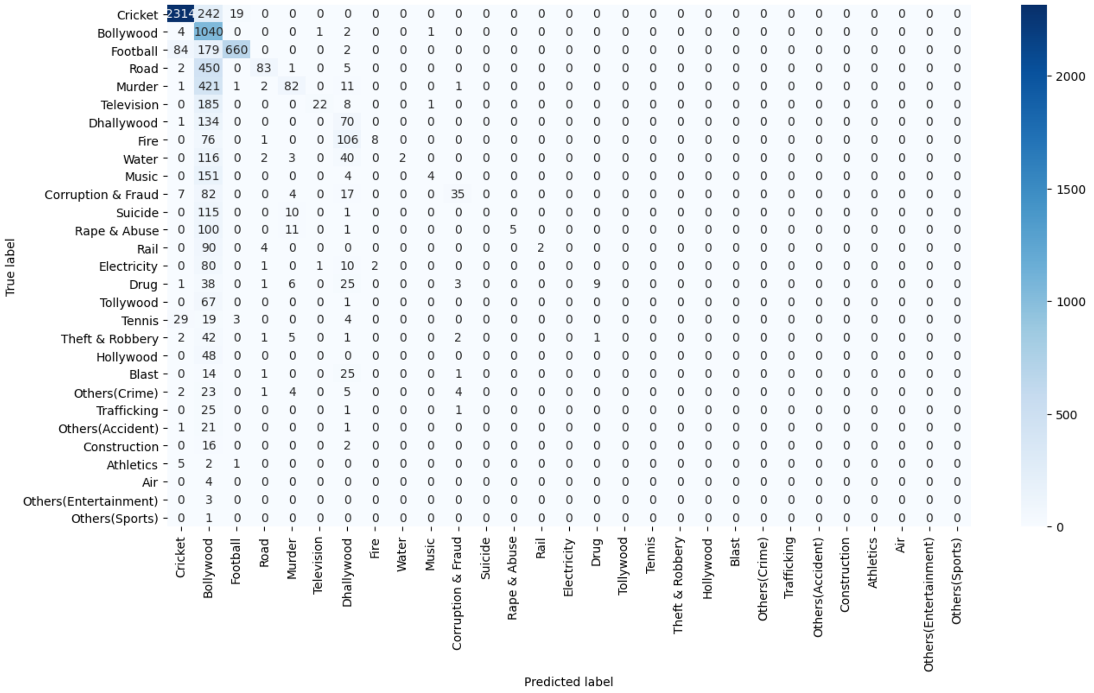


Figure 5.4: Confusion Matrix using news headlines(29 subcategories)

Figure 5.3 and 5.4 portrays the confusion matrix for news classifiers using headlines to classify into 4 categories and 29 subcategories respectively. As regard to the headlines classification, *Accident* and *Crime* data are mostly misclassified with each other where 42 *Accident* data are misclassified as *Crime* and 54 *Crime* data are misclassified as *Accident* Mention how much data? (Figure ??. Such as an *Accident* headline data **"দিয়া-রাজীবের মৃত্যুর মামলায় তিনজনের যাবজ্জীবন, দুজন খালাস"** (Three sentenced to life imprisonment, two acquitted in Dia-Rajiv death case) confuses the model to think it as *Crime* data. Again the *Crime* headline **"লাশ উদ্ধার"** (Body recovered) may confuse with the *Accident* class. Similarly in subcategory classification, Road(Accident) is mostly misclassified as Bollywood data due to the similarity in headlines for both classes.

From both the confusion matrices, we can presume that categories with more training data perform more accurately. In most cases news headlines are not well-structured and are mostly deprived of the inner context of the news data which causes performance degradation. Again news articles hold more words compared to the headlines which help the classifier model with more features to distinguish classes from one another.

For better insights, we investigate a few input for their actual class and predicted class by best three models including the proposed method. Table 5.9 shows the actual and predicted output. Called Avishek to write this table.

5.4 Comparison With Existing Techniques

Comparison with Existing Techniques is must (at least two recent methods on our dataset). Board members may not satisfied w/o comparison. Done

To evaluate the effectiveness of our model, we have compared our proposed system with existing approaches [11] [10]. The classifiers we have compared our system with are Logistic Regression (LR) with TF-IDF [11], Random Forest (RF) with TF-IDF [11] Naive Bayes (NB) with TF-IDF [11], Decision Tree (DT) with TF-IDF [10], K-Nearest Neighbour (KNN) with TF-IDF [10]and Support Vector Machine (SVM) with TF-IDF [10]. Table 5.10 summarizes the comparison performance for news classification into categories (4) and Table 5.11 shows the comparison performance for subcategory (29) classification. It can be seen that

Sample input	Actual class	**Predicted class**						
		MLP	LR	DT	RF	KNN	NB	SVM
গাজীপুরে ফ্যান কারখানায় অগ্নিকাণ্ডের ঘটনায় সোমবার রাতে জয়দেবপুর থানায় বিরুদ্ধে মামলা হয়েছে	Fire	Fire	Murder	Murder	Murder	Bollywood	Murder	Murder
ক্রিকেট বিশ্বকাপ আয়োজনের অভিজ্ঞতা বাংলাদেশের	Cricket	Cricket	Cricket	Cricket	Cricket	Bollywood	Cricket	Football
লালমনিরহাট রেলওয়ে স্টেশনে ট্রেনে কাটা পড়ে আমিনুল ইসলাম নামের এক পত্রিকা বিক্রেতা নিহত হয়েছেন	Rail	Rail	Murder	Rail	Murder	Bollywood	Murder	Road
চট্টগ্রামে পুলিশের ‘বন্দুকযুদ্ধে’ ছিনতাই মামলার এক আসামি নিহত হয়েছে	Theft & Robbery	Theft & Robbery	Murder	Murder	Murder	Bollywood	Murder	Murder
নেত্রকোনা আধুনিক সদর হাসপাতাল থেকে চুরি হওয়া নবজাতককে উদ্ধার	Theft & Robbery	Theft & Robbery	Murder	Theft & Robbery	Murder	Bollywood	Murder	Murder
দক্ষিণ কোরিয়ার পিয়ংইয়ংয়ে অনুষ্ঠিত এবারের শীতকালীন অলিম্পিকে অংশ নিয়েছিলেন ডানকান	Athletics	Football	Football	Football	Football	Bollywood	Football	Football

Table 5.9: Comparison among various baselines with the actual and predicted classes

our proposed system outperforms the machine learning approaches in terms of accuracy.

Method	Techniques	Ac (%)	
		News articles	News headlines
Alam et al. [11]	LR + TF-IDF	97.92	92.77
	RF + TF-IDF	97.22	90.61
	NB + TF-IDF	97.07	92.88
Mandal et al. [10]	DT + TF-IDF	93.77	88.31
	KNN + TF-IDF	54.08	59.68
	SVM + TF-IDF	98.01	93.82
Proposed	MLP + Keras embedding layer	**98.18**	**94.53**

Table 5.10: Performace comparison with other models for category(4) classification

Method	Techniques	Ac (%)	
		News articles	News headlines
Alam et al. [11]	LR + TF-IDF	89.20	78.64
	RF + TF-IDF	83.86	78.29
	NB + TF-IDF	70.18	70.00
Mandal et al. [10]	DT + TF-IDF	84.07	73.11
	KNN + TF-IDF	41.81	46.31
	SVM + TF-IDF	90.54	80.95
Proposed	MLP + Keras embedding layer	**90.63**	**82.97**

Table 5.11: Performace comparison with other models for subcategory(29) classification

5.5 Discussion

You should put the discussion. Critically analyze the results more better insights. Also mention your recommendation: What method is suitable for measuring semantic similarity in Bengali in the last paragraph. Done

From the comparative analysis of our proposed MLP classifier with the existing techniques, it can be seen that our system outperforms all of them. One of the reasons behind this is that MLP is a form of a neural network model that learns by calculating errors and updating them through backpropagation. Since machine learning models do not use backpropagation they don't have the privilege of learning from the errors and updating them. Another reason is that in neural networks the text data passes through several layers of interconnected nodes where each node matches the characteristics of the previous layer before passing it to other nodes. For these reasons our proposed classifier model performs better than the machine learning models.

5.6 Conclusion

In this chapter, we have presented the overall analysis of our system. It can be seen that our proposed method has outperformed the mentioned classifier models. The impact of our system from social and ethical perspectives is also visible. We have also discussed the overall errors and their context with examples giving an insight into our system faults.

Chapter 6

Conclusion

This Chapter summarizes the thesis with highlighting the major contributions of this work including a few weaknesses in the current implementation. This Chapter also provides the few recommendations for further improvements of the proposed system.

6.1 Conclusion

Summarizes your key contributions with mentioning the performance. Rewrite this section.

In this work, we have developed a corpus containing 76343 Bengali news articles with 523403 unique words and 29 classes which helped us in building the news classifier model. We illustrated and investigated various word embedding techniques including Keras Embedding Layer, Word2Vec and FastText with hyperparameters tuning for Bengali news classification. Performance analysis of hyperparameters for the MLP model has also been reported. We have developed an MLP-based method to classify Bengali news articles into 4 news classes and 29 sub news classes. Our MLP model with Keras Embedding layer has achieved an accuracy of 98.18% for category classification and an accuracy of 90.63% for subcategory classification when the news articles are taken into consideration. We have investigate and compared the performance of our proposed model with other ML baselines and existing techniques. Our MLP news classifier outperformed in both cases. To the best of our knowledge, this is the first work that has been done for classification into subcategories using both articles and headlines for Bangla news.

6.1.1 Limitations

Mention 4-5 limitations. Use bullet points Done

Our work is not above limitations. Some of our limitations are:

- Due to a shortage of time, we were not able to work with the whole news dataset we have collected. Using more data can help increasing the accuracy.
- There are several news categories such as International news that have not been taken into consideration.
- Even though we have achieved a high amount of accuracy in both category and subcategory classification but still there are some amount of errors that might be overcome with the latest sequence-to-sequence models and attention-based models.
- There are several news articles which has overlapped two or more categories. But multiclass classifcation system assigns only one label to each documents.

6.2 Future Recommendations

Mention 4-5 recommendations. Use bullet points Done As news classification into subcategories is the newest work in the field of Bangla news classification, there are various chances of improving our proposed system. Some of the future recommendations are:

- Neural Networks like CNN and RNN may perform better to investigated.
- Use state-of-art methods like attention-based models to make better predictions and to overcome the errors made by our system.
- Include more categories and subcategories along with more data.
- Topics in the subcategories are sometimes overlapped causing the model to get confused. To overcom this multilabel classification can be applied.

These issues are left to be worked on in the future.

References

[1] Wikipedia contributors, *Bengali language — Wikipedia, the free encyclopedia*, [Online; accessed 12-April-2021], 2021. [Online]. Available: `https://en.wikipedia.org/w/index.php?title=Bengali_language&oldid=1015708349` (cit. on p. 1).

[2] E. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text classification: A recent overview," p. 125, Jan. 2005 (cit. on p. 8).

[3] Satvika, V. Thada, and J. Singh, "A primer on word embedding," in *Data Intelligence and Cognitive Informatics*, I. Jeena Jacob, S. Kolandapalayam Shanmugam, S. Piramuthu, and P. Falkowski-Gilski, Eds., Singapore: Springer Singapore, 2021, pp. 525–541 (cit. on p. 10).

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, *Enriching word vectors with subword information*, 2017. arXiv: `1607.04606 [cs.CL]` (cit. on pp. 11, 25, 31).

[5] P. Marius, V. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, vol. 8, Jul. 2009 (cit. on p. 11).

[6] A. J. Stein, J. Weerasinghe, S. Mancoridis, and R. Greenstadt, "News article text classification and summary for authors and topics," *Computer Science Information Technology (CS IT)*, 2020. DOI: `10.5121/csit.2020.101401` (cit. on p. 13).

[7] R. Dutta, B. Jana, and M. Majumder, "Semantic similarity and word-net based web news classification," in *Intelligent Techniques and Applications in Science and Technology*, S. Dawn, V. E. Balas, A. Esposito, and S. Gope, Eds., Springer International Publishing, 2020, pp. 728–735 (cit. on p. 13).

[8] N. Isnaini, Adiwijaya, M. S. Mubarok, and M. Y. A. Bakar, "A multi-label classification on topics of indonesian news using k-nearest neighbor," *Journal of Physics: Conference Series*, vol. 1192, p. 012 027, Mar. 2019 (cit. on p. 13).

[9] D. B. Bracewell, J. Yan, F. Ren, and S. Kuroiwa, "Category classification and topic discovery of japanese and english news articles," *Electronic Notes in Theoretical Computer Science*, vol. 225, pp. 51–65, 2009, Proceedings of the Irish Conference on the Mathematical Foundations of Computer Science and Information Technology (MFCSIT 2006), ISSN: 1571-0661 (cit. on p. 13).

[10] A. K. Mandal and R. Sen, *Supervised learning methods for bangla web document categorization*, 2014. arXiv: `1410.2045 [cs.CL]` (cit. on pp. 14, 55–57).

[11] M. T. Alam and M. M. Islam, "Bard: Bangla article classification using a new comprehensive dataset," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018. DOI: 10.1109/icbslp.2018.8554382 (cit. on pp. 14, 55–57).

[12] A. N. Chy, H. Seddiqui, and S. Das, "Bangla news classification using naive bayes classifier," Mar. 2014 (cit. on p. 14).

[13] D. Cecchini and L. Na, "Chinese news classification," *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2018. DOI: 10.1109/bigcomp.2018.00125 (cit. on p. 14).

[14] B. Jang, I. Kim, and J. W. Kim, "Word2vec convolutional neural networks for classification of news articles and tweets," *Plos One*, vol. 14, no. 8, 2019 (cit. on p. 14).

[15] Z. Lu, W. Liu, Y. Zhou, X. Hu, and B. Wang, "An effective approach for chinese news headline classification based on multi-representation mixed model with attention and ensemble learning," in *Natural Language Processing and Chinese Computing*, Springer International Publishing, 2018, pp. 339–350 (cit. on p. 14).

[16] T. Alam, A. Khan, and F. Alam, "Bangla Text Classification using Transformers," 2020 (cit. on p. 14).

[17] M. M. Rahman, R. Sadik, and A. A. Biswas, "Bangla document classification using character level deep learning," *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 2020. DOI: 10.1109/ismsit50672.2020.9254416 (cit. on p. 15).

[18] M. Rabib, S. Sarkar, and M. Rahman, "Different Machine Learning based Approaches of Baseline and Deep Learning Models for Bengali News Categorization," *International Journal of Computer Applications*, vol. 176, no. 18, pp. 10–16, 2020 (cit. on p. 15).

[19] M. R. Hossain and M. M. Hoque, "Automatic bengali document categorization based on deep convolution nets," in *Emerging Research in Computing, Information, Communication and Applications*, N. R. Shetty, L. M. Patnaik, H. C. Nagaraj, P. N. Hamsavath, and N. Nalini, Eds., Singapore: Springer Singapore, 2019, pp. 513–525 (cit. on p. 15).

[20] R. Rahman, "A benchmark study on machine learning methods using several feature extraction techniques for news genre detection from bangla news articles titles," *7th International Conference on Networking, Systems and Security*, 2020. DOI: 10.1145/3428363.3428373 (cit. on p. 15).

[21] M. Shopon, "Bidirectional lstm with attention mechanism for automatic bangla news categorization in terms of news captions," *Lecture Notes in Electrical Engineering Electronic Systems and Intelligent Computing*, pp. 763–773, 2020. DOI: 10.1007/978-981-15-7031-5_72 (cit. on p. 15).

[22] R. Amin, N. S. Sworna, and N. Hossain, “Multiclass classification for bangla news tags with parallel cnn using word level data augmentation,” in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 174–177 (cit. on p. 15).

[23] M. M. H. Shahin, T. Ahmmed, S. H. Piyal, and M. Shopon, “Classification of bangla news articles using bidirectional long short term memory,” in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 1547–1551 (cit. on p. 15).

[24] *Prothom alo.* [Online]. Available: `http://www.daily-prothom-alo.com` (cit. on p. 18).

[25] *Daily nayadiganta.* [Online]. Available: `https://www.dailynayadiganta.com/` (cit. on p. 18).

[26] *Samakal.* [Online]. Available: `https://www.samakal.com/` (cit. on p. 18).

[27] *Kaler kantho.* [Online]. Available: `https://www.kalerkantho.com/` (cit. on p. 18).

[28] *Bhorer kagoj.* [Online]. Available: `https://www.bhorerkagoj.com/` (cit. on p. 18).

[29] J. K. and J. Saini, “Stop-word removal algorithm and its implementation for sanskrit language,” *International Journal of Computer Applications*, vol. 150, pp. 15–17, Sep. 2016 (cit. on p. 19).

[30] A. Rai and S. Borah, “Study of Various Methods for Tokenization,” in *App. of IoT*, Springer, pp. 193–200 (cit. on p. 30).

[31] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 281–305, 2012 (cit. on p. 33).