Bengali Documents categorization Using Deep Learning

BY

MD IBRAHIM KHOLIL
ID: 191-15-12766

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

Supervised By

Mst. Eshita Khatun
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Ms. Nusrat Jahan
Sr. Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2023

APPROVAL

This Project/internship titled **"Bengali Documents categorization Using Deep Learning"**, submitted by **Md Ibrahim Kholil**, ID No: **191-15-12766**, to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *23-01-2023*.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zahid Hasan — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Fahad Faisal — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Ahmed Wasif Reza — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
East West University

DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Mst. Eshita Khatun, Sr. Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Mst. Eshita Khatun
Sr. Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Ms. Nusrat Jahan
Sr. Lecturer
Department of CSE
Daffodil International University

Submitted by:

Md Ibrahim Kholil
ID: 191-15-12766
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

I really grateful and wish our profound my indebtedness to **Mst. Eshita Khatun**, **Sr. Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Research and development* " to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to Mst. Eshita Khatun, Nusrat Jahan and Dr Touhid Bhuiyan, Head**,** Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

In today's world, the amount of information and data is increasing day by day. After all, as the internet is readily available, a large amount of information and data is being stored in online cloud servers. Nowadays due to the availability of internet no one waits for newspapers anymore, everyone is more inclined towards online news portals. Online news portals are social media, Facebook, twitter, WhatsApp, telegram, Instagram, messenger, LinkedIn, blog etc. The amount of news on these online news portals is constantly increasing at a huge rate. As a result, the number of online readers is also increasing day by day. The need for data classification is increasing for all these digital data. There is a lot of data in the world that is not classified, such data is known as unusable data. Because we can use usable data but unusable data cannot be used for any purpose. The resulting data needs to be classified to become usable data. This research paper of ours is basically categorization of Bangla documents or data through deep learning. Our dataset had a total of 19137 data from which 18999 data were obtained by cleaning the data. Out of 18999 data, 13679 data have been taken for training, 1900 data for testing and 3420 data for validation check. The documents in our data set are divided into 12 categories, such as Politics, Education, Sports, Entertainment, Crime, Opinion, Accident, International, Environment, Economics, Science_Tech, Art. There are many types of deep learning models like CNN, LSTM, ANN, SBM etc. Among all the deep learning algorithms, the CNN model of our research paper is used. Using CNN model, we got 78.95% accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1. Introduction

The Internet is the main source of information and an integral part of people's lives. Online news sources are growing at a great rate in the world and due to the availability of internet, people are interested in reading daily news portals. Thousands of portals are providing hourly news updates and headlines in Bengali. In the digital age now most of the people read news through internet instead of reading news in newspapers. Nowadays online portals, Twitter, Facebook, blogs etc. are being used in a large amount through applications, due to which the use of internet is expanding and a large amount of information is available on websites. That is, the people of the world have become heavily dependent on Internet news. Also, in combination with high-speed Internet and handheld multimedia devices, users are creating and accessing large amounts of information every day. According to a report, there are about 80.83 million internet users at the end of January 2018, out of which about 30 million are social media users in Bangladesh. In recent years, the amount of online news production and access has increased day by day, with a focus on the Internet as the paper boom has been reduced. Many news organizations appear to be creating and uploading news online instead of releasing it. The news that is available from the Internet and the news that is published on the Internet is called e-news. This news readership is increasing day by day as a result of internet user scholarship. Due to which a large number of different news are being registered in the database of the website. News is a very important domain in developing countries like India, Bangladesh, Pakistan as news spreads knowledge as well as increases the level of public awareness about the news of neighboring countries. As a citizen of Bangladesh, Bengali language is our mother tongue and in 1952 many language martyrs gave their lives for the mother tongue Bengali. Bengali language is popular in several parts of India as well as Bangladesh. Currently, 228 million people speak Bengali in the world every day and 37 million foreigners speak Bengali. A count of speakers around the world shows that Bengali is the seventh most spoken language.

The last few decades have shown that both positive and negative effects of news reach the public very quickly. Many studies show that the negative impact of news has a negative effect on readers. Most of the videos seem to focus on releasing bad news rather than good news. These events or effects result in both physical and mental changes in individuals. Such psychological effects, such as negative thoughts or nervousness, depression, loss of concentration, stress, initiative, fear etc. Effective information retrieval is a principle of information technology. A higher generalization of text content is news headlines. The

internet caters to various types of news such as sports, computers, Hollywood, Bollywood, music, politics and social sciences. Users can locate and view any type of news from the internet. Through news headlines, users can easily search and view news according to their needs.

1.2. Motivation

Automatic text classification is an emerging topic. As the amount of digital data in the world is increasing at a massive rate, the need for text classification for this growing digital data is also increasing. Machine learning methods are used to classify data along with other data mining algorithms to classify text. Text classification is also used in some applications such as content tagging, spam filtering and business intelligence. Bengali language websites have huge amount of data from which it is quite difficult to find data. On the other hand, if you want to post on a forum, categorizing readers and hash tagging keywords becomes a very important task. Text categorization has not been done before in any application platform which is a big problem, thus status categorization has become very important for Bengali language. Again, the current need of text mining is increasing day by day due to which Bangladeshi and Indian researchers are focusing on creating applications using text mining. There are numerous works on text mining as well as sentiment analysis in Bengali which show results with good elimination. Most of the people read the news headlines before reading the news so that they can easily understand the sentiment of the news and what is wanted in the news. Any type of language problem in classification is said to be solved by NLP. As these problems express human language problem concepts and attempt to output them, machine learning algorithms are the most useful algorithms for understanding NLP problems. There are several categories of machine learning such as supervised, unsupervised and semi-supervised learning. Supervised, unsupervised and semi-supervised learning in supervised learning. Supervised learning needs to provide inputs and outputs as well as leveled data. Unsupervised learning requires providing input and output as well as level-free data. Semi-supervised learning is basically made up of supervised and unsupervised learning where labeled and unleveled data are combined. However, classifying so much data or news is time consuming, challenging and difficult. Machine learning algorithms are used to overcome these time consuming challenging and difficult algorithms. Text classification has been widely applied over the last few years with the development of machine learning. News headline classification is a type of text classification that can generally be divided into three parts, namely feature extraction, classifier extraction and evaluation. Another aspect of e-news in Bangladesh is that the readers prefer those sites which give regular breaking news and update news. The five popular sites are Pratham Alo, Bangladesh Pratid, Nayadigant, Jugantar, Samakal etc. Analyzing from Google trend data, it can be seen that the number of online readers of “Daily News” is decreasing day by day.

1.3. Relational of the Study

My research paper is to divide the news into several categories to find out which category this news falls into. In my research, news has been divided into 12 categories such as Politics, Education, Sports, Entertainment, Crime, Opinion, Accident, International, Environment, Economics, Science-Tech, Art. We have found several such works; they have also divided the news into different categories like ours and extracted the security using several algorithms.

1.4. Research Questions

We may have many types of questions in this study. For example-

1. What is the purpose of this research?
2. Why is this research being done?
3. Can we benefit from this research?
4. In which direction can we benefit from this research?
5. What are the results of this study?
6. What is the outcome of this study?

1.5. Expected Output

Since my research is to classify news headlines, once my research is complete, I can test with any type of data which class the news falls into. Besides, a lot of data can be saved from being destroyed, for example, there is a lot of unclassified data that is of no use, the unclassified data can be classified and converted into usable data.

1.6. Report Layout

However, due to improper use of data obtained from online, social media, i.e., not classifying, sorting and analyzing the data in a proper way, the news portals are failing to use a lot of potential data. If the data can be classified automatically using machine learning (ML), deep learning (DL) and NLP in a faster, more agile, cheaper and reliable way, this classification can be a huge potential solution. The remainder of the paper is divided into four sections: Section 2- background, Section 3- methodology, Section 4- experimental result and discussion, Section 5- impact on society, environment and sustainability. Finally, Section 6 will conclude the paper.

CHAPTER 2
BACKGROUND

2.1. Preliminaries

As the world is continually extending, occasions are going on all over. Also, the said incidents are spreading in online social media in a matter of moments because of web. Along these lines, the amount of news is expanding as time passes by. In any case, not to categorize the news, or at least, not to characterize which news falls under which classification. Because of this, a great deal of online information is lost which can't be utilized in future. So, to make every one of the information or archives usable, the information is classified utilizing various algorithms of machine learning and deep learning. So that the present information becomes usable for later.

2.2. Related Works

Prakash Kumar Sing et al. [1] propose a research paper which deep neural network model based on LSTM, SVM, HMM, CRF is used in this research paper. BiLSTM provided better accuracy 93.34% than other models. Mohammad Rabib Hossain et al. [2] propose a research paper different methods are used to classify news using machine learning baseline like SVM, Naive-Bayes, Random Forrest, Logistic Regression and deep learning model like BiLSTM, CNN of all these models. CNN provided the best accuracy 93.43%. Shazia Usmani et al. [3] propose a research paper, NLP based technique is used to classify Pakistani stock exchange news. Sharun Akter Khushbu et al. [4] propose a research paper is to extract the type of news from the headline of the news. For this, 5 machine learning classifiers such as SVM, NB, Logistic Regression, Neural Network, Random Forest have been used. Using the NN algorithm performed best, with an accuracy of 90%. Ruichao Wang et al. [5] propose a research paper, news headlines are detected by a hybrid system. These tests have been done through some systems like TFTrim, HybrideTrim, Topiary, TF, Hybrid, Trim, UTD etc. Through which TFTrim system has got the best results. Syeda Sumbul Hossain et al. [6] propose a research paper which 3383 news headlines were examined in this research paper. 7 machine learning models like Naïve Bayes 80.57%, Multinomial Naïve Bayes (MNB) 76.51%, Bernoulli Naïve Bayes 82.68%, Logistic regression 76.05%, Stochastic gradient descent (SGD) 74.55%, Linear support vector classifier (SVC) 75.90%, Nu support vector classifer 75.75% and two planning models such as Long short-term memory (LSTM) 68.82%, Convolutional neural network (CNN) 70.33% are used for sentiment analysis of news headlines. In machine learning Bernoulli Naïve Bayes and in deep learning Convolutional Neural Network (CNN) performed well. Paulo Santos et al. [7] propose a research paper, news headlines

are used to categorize. For this experiment SMO, Random Forest_1, Random Forest_2 classifier was used. In this experiment 62.50%, 57.50%, 61.00% accuracy was found in without relations as a feature and 62.70%, 59.00%, 63.50% accuracy was found respectively in with relations as a feature. Uchchhwas Saha et al. [8] propose a research paper analyzes the sentiment of Bengali comments using a hybrid approach, FirstText, Deep Learning classifier. "Adam", "Glove" is used in hybrid model and BiLSTM, CNN in deep learning. The hybrid (89.89%) model obtained higher accuracy than the FastText (62.25%) model. A.N.M. JuBaer et al. [9] propose a research paper presents several ways to categorize toxic comments. A few models like MultinomialNB 52.30%, SVM 30.76%, MultinomialNB (from scikit-learn) 52.30%, GausseanNB 49.23%, Classifier Chain with MultinomialNB 52.30%, Label Powerset with MultinomialNB 58.46%, MLkNN 58.46%, BP-MLL Neural Networks 60.00% have been used. Among all the models, the BPMLL Neural Network performed well. Mushfiqus Salehin et al. [10] propose a research paper attention mechanism-based sequence to sequence model is proposed. Worked here with own Bengali dataset and achieved good results from other research papers. Adrita Barua et al. [11] propose a research paper classifies NLP related language detection through machine learning. Here 6 machine learning algorithms such as Logistic Regression (LR), Support Vector Classifier (SVC), Decision Tree (DT), Multinomial Naive Bayes (MNB), Random Forest (RF), Term Frequency-Inverse Document Frequency (TF-IDF) are discussed in 4 categories (cricket, football, tennis and athletics). The highest accuracy is 97.60% obtained with the SVC algorithm. Md. Majedul Islam et al. [12] propose a research paper, which is detect Bengali word title sentiment using supervised algorithm. 5 Machine Learning Algorithms such as KNN, Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF) are discussed here. Highest accuracy is 75% obtained with SVM algorithm. Ettilla Mohiuddin Eumi et al. [13] propose a research paper focuses on classifying Bengali news headlines. Here the accuracy is calculated using some algorithms like LibSVM with stop words, LibSVM without stop words, Scikit Learn Librery, SVM, NB, LR, RF, Sequential deep learning, BiGRU Model. Headlines are divided into 6 different categories viz. The highest accuracy was 84% obtained using the BiGRU model algorithm. Md. Rafiuzzaman Bhuiyan et al. [14] propose a research paper focuses on categorizing news headlines. Here LSTM algorithm model is created and accuracy is calculated with 4580 trained data. Headlines are divided into 4 different categories like Rational, International, Science, Sports. Accuracy has come from 91.22% using LSTM model. Hoda Ahmed Galal Elsayed et al. [15] propose a research paper, the psychological influence of news headlines on readers is extracted through machine learning algorithm. Emotions are divided into 7 categories like anger, disgust, fear, happiness, neutral, sadness, and surprise. 6 machine learning classifiers like zeroR, KNN, CNN, decision trees, naïve Bayes, random forest and SVM have been used for this. Accuracy is achieved by using multilevel CNN 89.3%. Ronald Tudu et al. [16] propose a research paper classifies headlines with the help of machine learning. Headlines are divided into 10 different categories such as

Crime (CR) Economics (EC) International (IN) Sports (SP) Accident (AC) Environment (EV) Science and Technology (ST) Entertainment (EN) Politics (PO) Education (ED). Here the accuracy is measured using SVM, MNB, SGD, LR classifier. Accuracy is 87.5% achieved by using SVM algorithm. Fatema Jahara et al. [17] propose a research paper, Deep Learning Based Framework Multilayer Perception (MLP) classifier is used to classify newspaper headlines. Headlines are divided into 4 different categories such as accident, crime, entertainment, sports. Using the MLP algorithm, the highest accuracy was found for 98.18% (news articles) and 94.53% (news headlines). Raghad Bogery et al. [18] propose a research paper discusses some machine learning algorithms including NLP for classifying a large number of news headlines. 5 machine learning classifiers like KNN, SVM, NB, MNB and Gradient boosting have been used for this. The best performance was achieved using the Multinomial Naïve Bayes algorithm, with accuracy 90.12% and recall 90%. Headlines are divided into 3 different categories like travel, style & beauty, parenting. Md. Ferdouse Ahmed Foysal et al. [19] propose a research paper using LSTM algorithm of machine learning to classify news headlines achieved good accuracy. Accuracy is brought to 84% using LSTM algorithm. Headlines are divided into 5 different categories such as entertainment, national, sports, city state news. Ke Yahan et al. [20] propose a research paper takes a dataset containing 18 years of news and classifies it using machine learning NLP. 4 machine learning classifiers like Decision tree, Random Forest, SVC, NN have been used for this. The best performance was achieved using the NN algorithm, which has an accuracy of 0.8622.

2.3. Comparative Analysis and Summary

Table-1: Summary of Related Works

Paper No	Authors & Year	Used All Models	Height Accuracy Model	Height Model (%)
1	Prakash Kumar Sing (2021)	LSTM, SVM, HMM, CRF	LSTM	93.34%
2	Mohammad Rabib Hossain (2020)	ML- SVM, Naive-Bayes, Random Forrest, Logistic Regression DL- BiLSTM, CNN	CNN	93.43%
3	Shazia Usmani (2020)	NLP based technique	-	-

4	Sharun Akter Khushbu (2020)	SVM, NB, Logistic Regression, Neural Network, Random Forest	NN	90%
5	Ruichao Wang (2014)	TFTrim, HybrideTrim, Topiary, TF, Hybrid, Trim, UTD	TFTrim	Maximum Accuracy
6	Syeda Sumbul Hossain (2021)	ML- NB, Multinomial NB, Bernoulli NB, Logistic regression, Stochastic gradient descent (SGD), Linear support vector classifer (SVC), Nu support vector classifer. DL- LSTM, CNN.	ML- Bernoulli NB DL- CNN	ML- 82.68% DL- 70.33%
7	Paulo Santos (2015)	SMO, Random Forest_1, Ramdom forest_2 classifier	SMO (Without relations as features) Ramdom forest_2 (With relations as features)	62.50% 63.50%
8	Uchchhwas Saha (2022)	BiLSTM, CNN, FastText, Hybrid.	Hybrid	89.89%
9	A.N.M. JuBaer (2019)	Multinomial NB, SVM, Multinomial NB (from scikit-learn), Gaussean NB, Classifier Chain with Multinomial NB, Label Powerset with Multinomial NB, MLKNN, BP-MLL Neural Networks.	BPMLL Neural Network	60.00%

10	Mushfiqus Salehin (2019)	attention mechanism-based sequence-to-sequence model.	-	Best Accuracy
11	Adrita Barua (2021)	Logistic Regression (LR), Support Vector Classifer (SVC), Decision Tree (DT), Multinomial Naive Bayes (MNB), Random Forest (RF), Term Frequency-Inverse Document Frequency (TF-IDF)	SVC	97.60%
12	Md. Majedul Islam (2019)	KNN, Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF)	SVM	75%
13	Ettilla Mohiuddin Eumi (2021)	LibSVM with stop words, LibSVM without stop words, Scikit Learn Librery, SVM, NB, LR, RF, Sequential deep learning, BiGRU	BiGRU	84%
14	Md. Rafiuzzaman Bhuiyan (2021)	LSTM	LSTM	91.22%.
15	Hoda Ahmed Galal Elsayed (2020)	zeroR, KNN, Multilevel CNN, decision trees, Naïve Bayes, random forest and SVM	Multilevel CNN	89.3%.
16	Ronald Tudu (2018)	SVM, MNB, SGD, LR classifier	SVM	87.5%.

17	Fatema Jahara (2022)	deep learning-based framework multilayer perception (MLP) classifier	MLP (news articles) & MLP (news headlines)	98.18% & 94.53%
18	Raghad Bogery (2019)	KNN, SVM, NB, MNB and Gradient boosting	Multinomial Naïve Bayes	90.12%
19	Md. Ferdouse Ahmed Foysal (2021)	LSTM	LSTM	84%
20	Ke Yahan (2018)	Decision tree, Random Forest, SVC, NN	NN	86.22%

2.4. Scope of the Problem

Over the long time, various things are occurring on the world at various times. Once more, because of online they are accessible via social media, Facebook, twitter, WhatsApp, viber, Pinterest, telegram, messenger, google, YouTube and so forth. In this manner, the amount of news is expanding day by day. The expansion in how much news is the expansion in how much information or records on the online. In any case, not to arrange the said information or data, that is to say, not to characterize which information falls under which classification. Because of this, a great deal of online information is lost which can't be utilized in future. In this manner, to make this large number of information or reports usable, different algorithms of machine learning and deep learning, for example, logistic regression, linear regression, random forest classifier, decision tree, naive biyas, SVM, CNN, RNN, ANN, LSTM and so on the information is classified. Accordingly, the data or information got online can be utilized for different purposes.

2.5. Challenges

In this day and age how, much news is expanding day by day, in a couple of days there will be more unusable information than usable information on the online. The justification for the un-convenience of the said information or reports isn't to classify, that is to say, not to characterize which information falls under which classification. Because of this, a lot of online information is lost which can't be utilized in future. In the event that we don't change the usability of this multitude of information now, we will deal with numerous issues from now on. Involving information in the future to make changes will call for a ton of investment and talented labor. This will burn through both time and money. In this way, assuming the information got now for example the information accessible via social media or different stages is classified then the information will be useful and won't deal with issues from now on.

CHAPTER 3
RESEARCH METHODOLOGY

3.1. Design Approach

Supervised learning is used in our data set for multivariate classification problems. This dataset is used in deep learning such as CNN. Before applying this model, the data set is pre-processed, then the data is cleaned, then tokenization is done by dividing the entire data set into small tokens. After data pre-processing and cleaning, the data set is divided into three parts, i.e., training set, testing set and validation set. Then the algorithm is applied on the training and testing set, then the performance of the algorithm is tested and the results are obtained. The architecture diagram for the entire system is shown in figure:

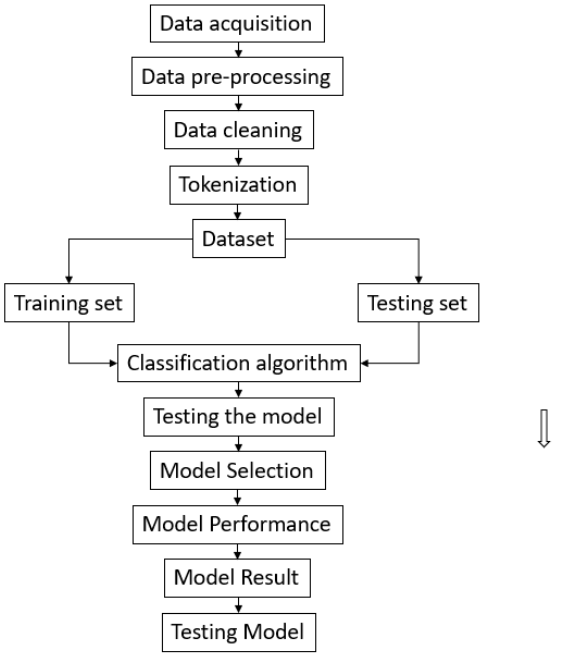

Figure-1: Architecture of Working Process

3.2. Dataset Description

The main dataset used in this paper consists of 3 columns namely label, text and is_valid. The category is divided into 12 categories, namely- Politics, Education, Sports, Entertainment, Crime, Opinion, Accident, International, Environment, Economics, Science-Tech and Art. To run the machine learning and deep

learning algorithms, we used a basic partitioning of the data set consisting of 13679 samples for training, 1900 samples for testing, and 3420 samples for validation.

Table-2: Category name or Description

Source News	Category	Description	Label
Online News Headline	Politics	Politics news will be shown from all types of news.	Politics, Education, Sports, Entertainment, Crime, Opinion, Accident, International, Environment, Economics, Science_Tech, Art
	Education	Education news will be shown from all types of news.	
	Sports	Sports news will be shown from all types of news.	
	Entertainment	Entertainment news will be shown from all types of news.	
	Crime	Crime news will be shown from all types of news.	
	Opinion	Opinion news will be shown from all types of news.	
	Accident	Accident news will be shown from all types of news.	
	International	International news will be shown from all types of news.	
	Environment	Environment news will be shown from all types of news.	
	Economics	Economics news will be shown from all types of news.	
	Science_Tech	Science_Tech news will be shown from all types of news.	
	Art	Art news will be shown from all types of news.	

3.3. Dataset Utilized

3.3.1. Pre-Processing

The data we usually get from online is basically unstructured data. So, after data collection all unstructured data is converted into structured data through data pre-processing. Because after collecting data from online social media, the data contains a lot of duplicate data, null value, bad data etc. Dataset is processed to make sense, remove duplicate data, convert to unique data, remove short length data. Consequently, it is necessary to apply the Pak processing method to the dataset. Also, dataset partitioning, unique words of each class, number of documents in each class, word list, dictionary sorting of word list, total words of each class etc. were extracted from the data set. To apply pre-processing method, I follow below steps-

- Remove URLs, screen names and hashtags.
- Remove zero values, emojis, symbols, punctuation and numbers.
- Remove all retweets and unnecessary symbols.
- Remove all low length data.

First prepare the dataset and distribute the dataset. A total of 19137 documents appears in the dataset. The 19137 documents are divided into 12 categories (Politics, Education, Sports, Entertainment, Crime, Opinion, Accident, International, Environment, Economics, Science Tech, Art), among which politics documents are the most, followed by sports and education almost equally so these two categories are ranked second highest, followed by entertainment third highest, followed by Crime, Opinion, Accident, International, Environment, Economics, Science Tech and finally Art related documents respectively have less. Below is a graph of the data distribution:

Figure-2: Plot the class distribution

3.3.2. Dataset Cleaning

The dataset has total 12 News Categories data and politics class has maximum number of articles. All types of emojis, symbols, punctuation, punctuation from new 12 categories of data set like- (()‚ {}‚ []‚ /‚ " "‚ ' '‚ :‚ !‚ ?) cleaning has been done by removing etc. Also, cleaning was done by identifying short length documents and removing them from the data set. 138 small documents were removed and remaining documents were 18999.

Table-3: Cleaned data from dataset

No	Original Text	Label	Cleaned Text
1	ঢাকা, সেপ্টেম্বর ২৭ (বিডিনিউজ টোয়েন্টিফোর ডটকম)- রাজধানীর আজিমপুরে বাবার কবরে চিরনিদ্রায় শায়িত হলেন বিশিষ্ট সাংবাদিক আতাউস সামাদ, ১৯৮০ এর দশকে এরশাদবিরোধী আন্দোলনের সময় বস্তুনিষ্ঠ সংবাদ পেতে দেশবাসী বিবিসিতে যার কণ্ঠ শোনার অপেক্ষায় থাকত।	Politics	ঢাকা, সেপ্টেম্বর ২৭ বিডিনিউজ টোয়েন্টিফোর ডটকম রাজধানীর আজিমপুরে বাবার কবরে চিরনিদ্রায় শায়িত হলেন বিশিষ্ট সাংবাদিক আতাউস সামাদ ১৯৮০ এর দশকে এরশাদবিরোধী আন্দোলনের সময় বস্তুনিষ্ঠ সংবাদ পেতে দেশবাসী বিবিসিতে যার কণ্ঠ শোনার অপেক্ষায় থাকত
2	বিএড সনদ অর্জনকারীদের উচ্চতর স্কেল প্রদানের ক্ষেত্রে শিক্ষা মন্ত্রণালয়ের যুগ্ম সচিব ড. ফারুক হোসেন স্বাক্ষরিত ২০১৬ খ্রিষ্টাব্দের ২৮ জানুয়ারি জারি করা পত্র অনুসরণ করার জন্য মাধ্যমিক ও উচ্চ শিক্ষা অধিদপ্তরকে নির্দেশ দেওয়া হয়েছে।	Education	বিএড সনদ অর্জনকারীদের উচ্চতর স্কেল প্রদানের ক্ষেত্রে শিক্ষা মন্ত্রণালয়ের যুগ্ম সচিব ড ফারুক হোসেন স্বাক্ষরিত ২০১৬ খ্রিষ্টাব্দের ২৮ জানুয়ারি জারি করা পত্র অনুসরণ করার জন্য মাধ্যমিক ও উচ্চ শিক্ষা অধিদপ্তরকে নির্দেশ দেওয়া হয়েছে
3	ঢাকা, অক্টোবর ০৮ (বিডিনিউজ টোয়েন্টিফোর ডটকম)- ম্যাচ গড়পেটায় "জড়িত" থাকার অভিযোগের তদন্ত শেষ না হওয়া পর্যন্ত ৬ আম্পায়ারকে সব ধরনের ক্রিকেট ম্যাচে সাময়িকভাবে নিষিদ্ধ করেছে আইসিসি।	Sports	ঢাকা, অক্টোবর ০৮ বিডিনিউজ টোয়েন্টিফোর ডটকম ম্যাচ গড়পেটায় জড়িত থাকার অভিযোগের তদন্ত শেষ না হওয়া পর্যন্ত ৬ আম্পায়ারকে সব ধরনের ক্রিকেট ম্যাচে সাময়িকভাবে নিষিদ্ধ করেছে আইসিসি
4	শুক্রবার রাতে-- উৎসবের ফেইসবুক পেইজে এ ঘোষণা দিল…	Entertainment	শুক্রবার রাতে উৎসবের ফেইসবুক পেইজে এ ঘোষণা দিল…

	এ যেন একেবারে হাতে লাল টিসি ধরিয়ে দিয়ে বলা “নিষ্ক্রান্ত”… গ্লিটজকে শাবনূর বললেন, “নিন্দুকেরা নানা কথা রটিয়েছে কানে এসে…		এ যেন একেবারে হাতে লাল টিসি ধরিয়ে দিয়ে বলা নিষ্ক্রান্ত… গ্লিটজকে শাবনূর বললেন নিন্দুকেরা নানা কথা রটিয়েছে কানে এসে…
5	নাটোরে চলন্ত বাস থেকে এক যাত্রীকে ধাক্কা দিয়ে ফেলে দেওয়ার অভিযোগ উঠেছে এসবি পরিবহনের এক বাস তত্ত্বাবধায়কের বিরুদ্ধে। লালপুর উপজেলার গড়মাটি এলাকায় ঢাকা-কুষ্টিয়া সড়কে গতকাল মঙ্গলবার এ ঘটনা ঘটে। আহত যাত্রীর নাম আহসান হাবীব।	Crime	নাটোরে চলন্ত বাস থেকে এক যাত্রীকে ধাক্কা দিয়ে ফেলে দেওয়ার অভিযোগ উঠেছে এসবি পরিবহনের এক বাস তত্ত্বাবধায়কের বিরুদ্ধে লালপুর উপজেলার গড়মাটি এলাকায় ঢাকা কুষ্টিয়া সড়কে গতকাল মঙ্গলবার এ ঘটনা ঘটে আহত যাত্রীর নাম আহসান হাবীব
6	আজ থেকে অবশ্যই শেখ হাসিনার নেতৃত্বাধীন সরকার “অবৈধ” হবে না। কিন্তু আজকের দিনটি সাংবিধানিকভাবে বিশেষ তাৎপর্যমণ্ডিত।	Opinion	আজ থেকে অবশ্যই শেখ হাসিনার নেতৃত্বাধীন সরকার অবৈধ হবে না কিন্তু আজকের দিনটি সাংবিধানিকভাবে বিশেষ তাৎপর্যমণ্ডিত
7	রাজধানীর নিমতলী, চুড়িহাট্টা, কামালবাগসহ পুরান ঢাকার বিভিন্ন স্থানে ঝুঁকিপূর্ণ কেমিক্যাল ব্যবসার কারণে সৃষ্ট অগ্নিকাণ্ডে গত এক দশকে জীবনপ্রদীপ নিভে গেছে দুই শতাধিক মানুষের।	Accident	রাজধানীর নিমতলী চুড়িহাট্টা কামালবাগসহ পুরান ঢাকার বিভিন্ন স্থানে ঝুঁকিপূর্ণ কেমিক্যাল ব্যবসার কারণে সৃষ্ট অগ্নিকাণ্ডে গত এক দশকে জীবনপ্রদীপ নিভে গেছে দুই শতাধিক মানুষের
8	নওয়াজ শরীফ এই আন্দোলনকে‘ছোটখাটো ঝাপটা’ বলে উড়িয়ে দিয়েছেন। তিনি বলেছেন, এই আন্দোলনের মুখে তাঁর পদত্যাগের প্রশ্নই আসে না। এনডিটিভি বলেছে, আজ রাতে পার্লামেন্ট ভবন এলাকা থেকে আন্দোলনকারীরা প্রধানমন্ত্রীর বাসভবনের দিকে অগ্রসর হওয়া শুরু করে।	International	নওয়াজ শরীফ এই আন্দোলনকে ছোটখাটো ঝাপটা বলে উড়িয়ে দিয়েছেন তিনি বলেছেন এই আন্দোলনের মুখে তাঁর পদত্যাগের প্রশ্নই আসে না এনডিটিভি বলেছে আজ রাতে পার্লামেন্ট ভবন এলাকা থেকে আন্দোলনকারীরা প্রধানমন্ত্রীর বাসভবনের দিকে অগ্রসর হওয়া শুরু করে

9	জেরুজালেম, মে ০২ (বিডিনিউজ টোয়েন্টিফোর ডটকম/রয়টার্স)- রকেট হামলা ও অস্ত্র চোরাচালান বন্ধ হলে গাজা উপত্যকায় হামাসের যুদ্ধবিরতির প্রস্তাব মেনে নিতে পারে ইসরায়েল। তবে কোনো আনুষ্ঠানিক চুক্তি ছাড়াই নীরবে তা করা হতে পারে।	Environment	জেরুজালেম মে ০২ বিডিনিউজ টোয়েন্টিফোর ডটকম রয়টার্স রকেট হামলা ও অস্ত্র চোরাচালান বন্ধ হলে গাজা উপত্যকায় হামাসের যুদ্ধবিরতির প্রস্তাব মেনে নিতে পারে ইসরায়েল তবে কোনো আনুষ্ঠানিক চুক্তি ছাড়াই নীরবে তা করা হতে পারে
10	গতকাল সোমবার বাংলাদেশ উন্নয়ন গবেষণা প্রতিষ্ঠানের (বিআইডিএস) এক সেমিনারে অর্থনীতিবিদেরা এ অভিমত তুলে ধরেন। রাজধানীর আগারগাঁওয়ে বিআইডিএসের সম্মেলনকক্ষে বিআইডিএস ও জাপান এক্সটারনাল ট্রেড অর্গানাইজেশন (জেইটিআরও) যৌথভাবে "পূর্বমূখী নীতি: বাংলাদেশ প্রেক্ষিত" শীর্ষক এই সেমিনারের আয়োজন করে।	Economics	গতকাল সোমবার বাংলাদেশ উন্নয়ন গবেষণা প্রতিষ্ঠানের বিআইডিএস এক সেমিনারে অর্থনীতিবিদেরা এ অভিমত তুলে ধরেন রাজধানীর আগারগাঁওয়ে বিআইডিএসের সম্মেলনকক্ষে বিআইডিএস ও জাপান এক্সটারনাল ট্রেড অর্গানাইজেশন জেইটিআরও যৌথভাবে পূর্বমূখী নীতি বাংলাদেশ প্রেক্ষিত শীর্ষক এই সেমিনারের আয়োজন করে
11	ইএটিএল-প্রথম আলো অ্যাপস প্রতিযোগিতা ২০১৫ বিশ্ববিদ্যালয় কার্যক্রম শুরু হয়েছে। এরই অংশ হিসেবে গত বৃহস্পতিবার ঢাকার ড্যাফোডিল আন্তর্জাতিক বিশ্ববিদ্যালয়ে (ডিআইইউ) স্ট্যাট বিশ্ববিদ্যালয় ও ইন্টারন্যাশনাল ইউনিভার্সিটি অব বিজনেস অ্যাগ্রিকালচার অ্যান্ড টেকনোলজিতে শুরু হলো বিশ্ববিদ্যালয় কার্যক্রম।	Sciencs_tech	ইএটিএল প্রথম আলো অ্যাপস প্রতিযোগিতা ২০১৫ বিশ্ববিদ্যালয় কার্যক্রম শুরু হয়েছে এরই অংশ হিসেবে গত বৃহস্পতিবার ঢাকার ড্যাফোডিল আন্তর্জাতিক বিশ্ববিদ্যালয়ে ডিআইইউ স্ট্যাট বিশ্ববিদ্যালয় ও ইন্টারন্যাশনাল ইউনিভার্সিটি অব বিজনেস অ্যাগ্রিকালচার অ্যান্ড টেকনোলজিতে শুরু হলো বিশ্ববিদ্যালয় কার্যক্রম
12	নানা রং দিয়ে তৈরি একটি পূর্ণাবয়ব মূর্তি। তাকে আধার করে লাল, হলুদ, কালো, আর গোলাপি রঙের মিশেলে তৈরি অপূর্ব একটি কারুকাজ। সঙ্গে সূর্যমুখীর পাপড়ির মতো তিনটি ফুল ফুটে রয়েছে।	Art	নানা রং দিয়ে তৈরি একটি পূর্ণাবয়ব মূর্তি তাকে আধার করে লাল হলুদ কালো আর গোলাপি রঙের মিশেলে তৈরি অপূর্ব একটি কারুকাজ সঙ্গে সূর্যমুখীর পাপড়ির মতো তিনটি ফুল ফুটে রয়েছে

After cleaning the data set i.e., removing different types of symbols, punctuation, emoji from the data set and removing 138 short documents, the maximum document length is 7405, the minimum document length is 21 and the average document length is 226. Below is the visualization after cleaning the dataset:

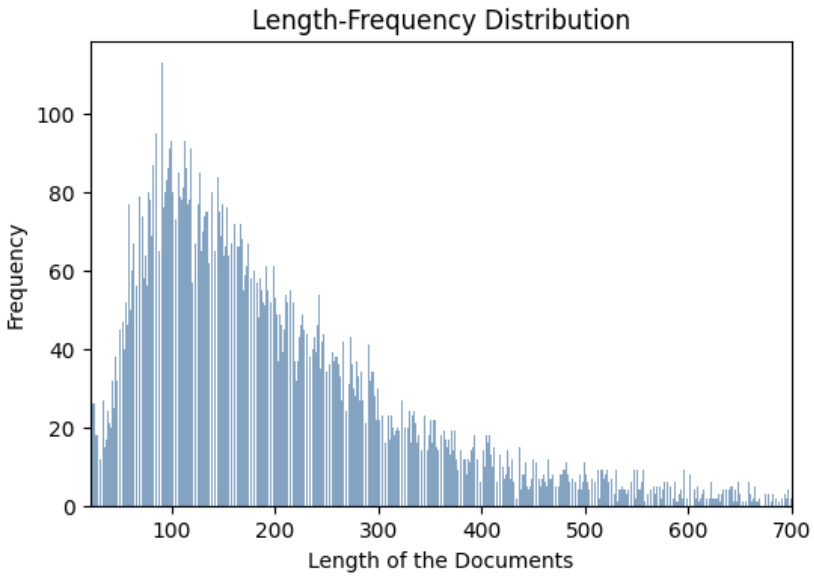

Figure-3: Length-Frequency Distribution

3.3.3. After pre-processing dataset summary

The total number of documents in the data set is 19137 and there are 3 columns. The data set is then divided into twelve new categories. The categories are - Politics, Education, Sports, Entertainment, Crime, Opinion, Accident, International, Environment, Economics, Science_Tech, Art. Within each category, the number of documents, number of words, number of unique words and most frequent words are extracted.

Class Name is politics, number of documents is 4012, number of words is 950344, number of unique words is 52967 and here are some picks on the words that have been used the most times. Most Frequent Words: শেখ 7014, প্রধানমন্ত্রী 6131, খালেদা 5189, হয়েছে 4753, হাসিনা 4611, আওয়ামী 4355, বিএনপি 4240, জানান 4114, বিডিনিউজ 4080, কথা 4012.

Class name is education, number of documents is 2368, number of words is 538268, number of unique words is 50415 and here are some picks on the words that have been used the most times. Most Frequent Words: ঘ 5203, ক 4680, খ 4677, গ 4648, শিক্ষা 3570, হয়েছে 3287, বিডিনিউজ 2784, টোয়েন্টিফোর 2603, ঢাকা 2432, ডটকম 2104.

Class name is sports, number of documents is 2360, number of words is 499145, number of unique words is 46268 and here are some picks on the words that have been used the most times. Most Frequent Words: খেলা 3047, রান 2927, বিডিনিউজ 2477, টোয়েন্টিফোর 2315, ডটকম 2288, ০০ 2122, দলের 2080, শেষ 2001, বাংলাদেশ 1890, এক 1845.

Class name is entertainment, number of documents is 2012, number of words is 426691, number of unique words is 51218 and here are some picks on the words that have been used the most times. Most Frequent Words: ০০ 5522, ৩০ 3658, সংবাদ 3302, সকাল 2195, ১০ 2193, রাত 2175, নাটক 2118, গান 2066, দুপুর 1695, ১২ 1693.

Class name is crime, number of documents is 1718, number of words is 305681, number of unique words is 35478 and here are some picks on the words that have been used the most times. Most Frequent Words: পুলিশ 2346, জানান 2014, এক 1775, হয়েছে 1556, বিডিনিউজ 1476, টোয়েন্টিফোর 1382, গ্রেপ্তার 1112, লাশ 1106, হত্যা 1103, থানার 1101.

Class name is opinion, number of documents is 1619, number of words is 582323 number of unique words is 63044 and here are some picks on the words that have been used the most times. Most Frequent Words: হয়েছে 2369, কথা 2364, এক 2286, সরকার 2010, নির্বাচন 1900, দেশের 1772, সরকারের 1649, রাজনৈতিক 1638, দলের 1376, হিসেবে 1373.

Class name is accident, number of documents is 1264, number of words is 180597, number of unique words is 24193 and here are some picks on the words that have been used the most times. Most Frequent Words: জানান 1735, বিডিনিউজ 1720, হয়েছে 1664, টোয়েন্টিফোর 1610, নিহত 1332, দুর্ঘটনা 1210, ডটকম 1157, আহত 1111, পুলিশ 1051, এক 981.

Class name is international, number of documents is 1015, number of words is 158806 is number of unique words is 25882 and here are some picks on the words that have been used the most times. Most Frequent Words: এক 1069, গত 810, হয়েছে 794, প্রেসিডেন্ট 615, কথা 531, গতকাল 503, বছর 423, মার্কিন 410, বিরুদ্ধে 398, রাশিয়া 393.

Class name is environment, number of documents is 830, number of words is 206180, number of unique words is 28824 and here are some picks on the words that have been used the most times. Most Frequent Words: পরিবেশ 1607, হয়েছে 1314, বিডিনিউজ 1090, টোয়েন্টিফোর 952, জানান 941, এক 880, ডটকম 821, ঢাকা 598, সরকার 584, গত 578.

Class name is economics, number of documents is 688, number of words is 134761, number of unique words is 18055 and here are some picks on the words that have been used the most times. Most Frequent Words: লেনদেন 1275, টাকা 1075, হয়েছে 1011, সূচক 861, দাম 825, শতাংশ 823, পয়েন্ট 794, গত 717, দশমিক 694, টাকার 647.

Class name is science_tech, number of documents is 581, number of words is 109353, number of unique words is 21866 and here are some picks on the words that have been used the most times. Most Frequent Words: ০০ 717, এক 550, সংবাদ 532, তথ্য 440, ৩০ 440, ১০ 366, টাকা 365, তৈরি 347, হয়েছে 330, অ্যাপল 292.

Class name is art, number of documents is 532, number of words is 192393, number of unique words:35588 and here are some picks on the words that have been used the most times. Most Frequent Words:০০ 2074, সংবাদ 1355, ৩০ 1242, রাত 1216, নাটক 862, এক 852, সকাল 841, ১০ 828, কথা 780, ধারাবাহিক 674.

After pre-processing the data, the most frequent words were calculated. Besides, how many total documents are there in each class, how many total words are there, how many unique words are there, a table is made with the number of these and it is also shown through visualization under the table.

Table-4: Dataset summary after pre-processing

No	Total Documents	Total Words	Unique Words	Class Names
0	4012	950344	52967	politics
1	2368	538268	50415	education
2	2360	499145	46268	sports
3	2012	426691	51218	entertainment
4	1718	305681	35478	crime
5	1619	582323	63044	opinion

6	1264	180597	24193	accident
7	1015	158806	25882	international
8	830	206180	28824	environment
9	688	134761	18055	economics
10	581	109353	21866	science_tech
11	532	192393	35588	art

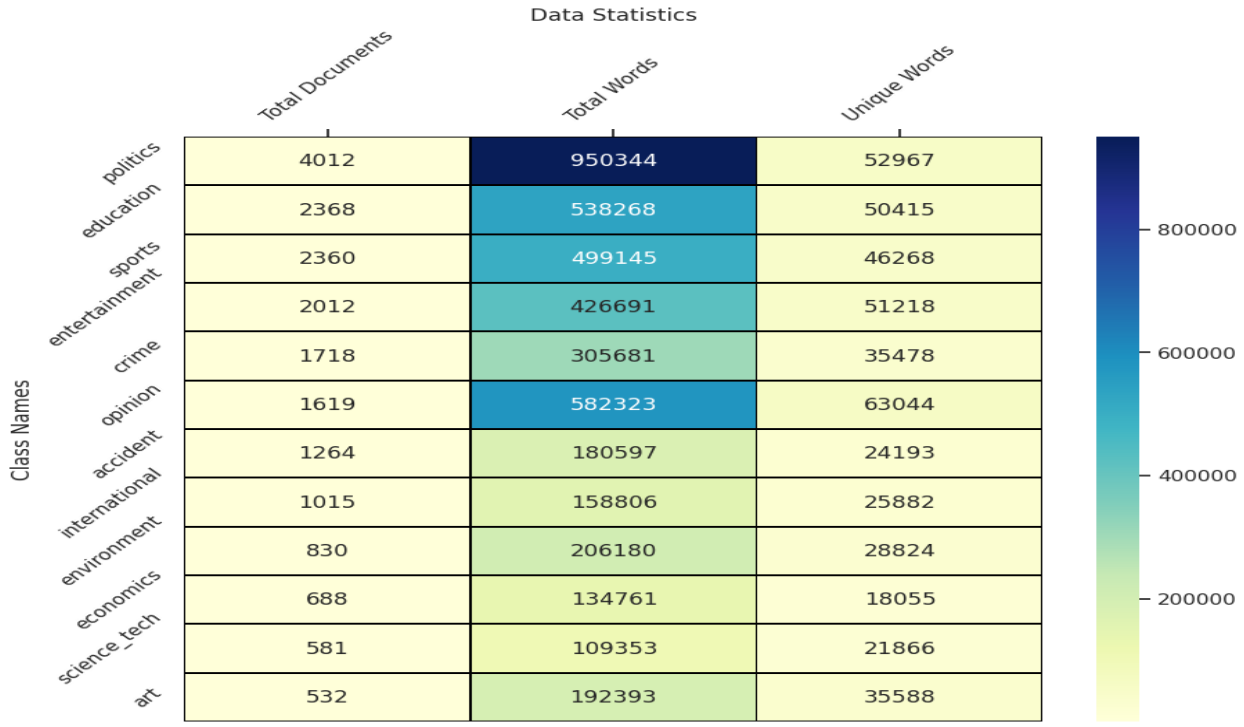

Figure-4: Data Statistics of dataset summary

3.3.4. Label Encoding

Level encoding is a method of converting character type data into numeric index data, so that they can be converted into a machine-readable form. When we work with different algorithms of machine algorithms, machines do not understand categorical data well or the results do not come well, so level encoding is used to solve this problem. This is an important step for data sets in supervised learning. Our dataset is divided into 12 categories with level encoding for each category.

Class Names: → [‘accident’ ‘art’ ‘crime’ ‘economics’ ‘education’ ‘entertainment’ ‘environment’ ‘international’ ‘opinion’ ‘politics’ ‘science tech’ ‘sports’].

After converting all the data into numerical language, the dataset is split. The dataset is divided into 2 parts, a) training data, b) test data. Out of total 18999 data, training data is 13679 and testing data is 1900. Some data is kept for validation checking, there are 3420 data for validation checking. Total parameters (933622), trainable parameters (933622) and non-trainable parameters (0) are fitted to the dataset using this model.

3.3.5. Tokenizer

Tokenization is the splitting of a sentence or paragraph or an entire text document into smaller parts. Tokenization is used a lot when working with text data. Our dataset is also tokenized. The total documents in our data set were 18999, from which 188977 unique tokens were obtained. Two types of tokenization are performed on our data set. Namely- Encoded Sequence and Padded Sequence.

i. Encoded Sequences-

Encoded Sequences
রোজা রাখার উদ্দেশ্যে শেষ রাতে উষা উদয়ের পানাহার সেহরি হিসেবে পরিচিত সেহরি উর্দু শব্দ মূল আরবি সুহুর শাব্দিক অর্থ নিদ্রাভঙ্গ নিদ্রাভঙ্গ ঘুম জেগে ওঠা রাত্রি জাগরণ রোজা পালনের সুবহে
[4970, 509, 1230, 41, 180, 1, 1, 1, 1, 27, 987, 1, 1, 1282, 352, 4287, 1, 1, 244, 1, 3665, 4126, 2245,1, 1, 4970, 1915, 1, 1, 1, 1, 1,1, 1, 1, 1, 1, 4691, 4307, 4970, 1915, 1, 1, 1, 3098, 1,1, 546, 4054, 374, 41, …]

ii. Paded Sequences-

Paded Sequences

রোজা রাখার উদ্দেশ্যে শেষ রাতে উষা উদয়ের পানাহার সেহরি হিসেবে পরিচিত সেহরি উর্দু শব্দ মূল আরবি সুহুর শাব্দিক অর্থ নিদ্রাভঙ্গ নিদ্রাভঙ্গ ঘুম জেগে ওঠা রাত্রি জাগরণ রোজা পালনের সুবহে
[1959 626 1 81 2327 599 1 1 1 1 1214 41 180 3665 1 1 2566 1 2908 4970 3899 1 1 68 4970 1975 1 1 1 347 68 1 2566 1 1 1 347 825 1 1 1 1028 822 1 1818 1 28 1 1 1 1 1 1 21 1 1982 1 1 1 227 3245 1 1 2 1 1 1 1 1 1 …]

3.4. Statistical Analysis

- The dataset contains a total of 18999 amount of data.
- The dataset keeps 3 columns.
- The dataset contains a total of 13679 training data.
- The dataset contains a total of 1900 testing data.
- The dataset contains a total of 3420 validation data.
- Total parameters (933622).
- Trainable parameters (933622).
- Non-Trainable parameters (0).
- Category are classified into 12 steps (Politics, Education, Sports, Entertainment, Crime, Opinion, Accident, International, Environment, Economics, Science_Tech, Art).

3.5. Convolutional Neural Network (CNN)

CNN Algorithm stands for Convolutional Neural Network also known as ConvNet. CNN stands for Supervised Learning, a subtype of neural network that specializes in data processing. CNN has five levels. Namely: Convolutional Layer, Pooling Layer, Fully Connected Layer, Dropout, Activation Functions. Its main advantage is that it detects important features automatically without any human supervision. There are many other types of neural networks in deep learning but CNN is the network architecture of choice for identifying and recognizing objects. CMM's built-in convolutional layer reduces image high-dimensionality

without losing information. This is why CNN is used in all cases. Examples of CNN are face recognition, image classification, speech recognition etc.

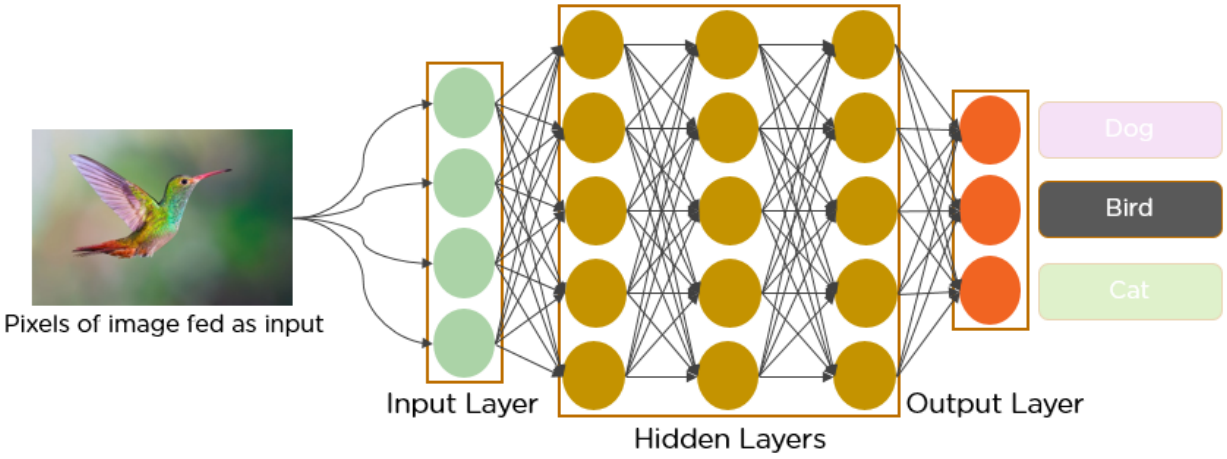

Figure-5: Visual overview of CNN

3.5.1. Validation and Training accuracy:

Validation accuracy and training accuracy were calculated with 10 epochs of the model. Validation Accuracy is found to be 0.5924 in epoch-1, accuracy is found to be 0.7243 in epoch-2, 0.7547 in epoch-3, epoch-4 the accuracy is found to be 0.7670, epoch-5 the accuracy is found to be 0.7675, accuracy is found to be 0.7588 in epoch-6, 0.7646 in epoch-7, accuracy is found to be 0.7558 in epoch-8, accuracy is found to be 0.7605 in epoch-9 and at epoch-10 the accuracy is found to be 0.7582. That is, it is seen here that the validation accuracy value is upward.

On the other hand, in training accuracy, the training accuracy in epoch-1 is 0.4190, in epoch-2 the training accuracy is 0.7082, in epoch-3 the training accuracy is 0.8216, in epoch-4 the training accuracy is 0.8699, training assurance in epoch-5 is 0.9005, training assurance in epoch-6 is 0.9226, training assurance in epoch-7 is 0.9349, in epoch-8 the training accuracy is 0.9448, in epoch-9 the training accuracy is 0.9550, training accuracy in epoch-10 is 0.9604. That is, the value of training loss is downward.

Table-5: Validation and Training accuracy table for CNN

Epoch No	Validation Accuracy	Training Accuracy
Epoch 1	0.5924	0.4190
Epoch 2	0.7243	0.7082

Epoch 3	0.7547	0.8216
Epoch 4	0.7670.	0.8699
Epoch 5	0.7675	0.9005
Epoch 6	0.7588	0.9226
Epoch 7	0.7646	0.9349
Epoch 8	0.7558	0.9448
Epoch 9	0.7605	0.9550
Epoch 10	0.7582.	0.9604

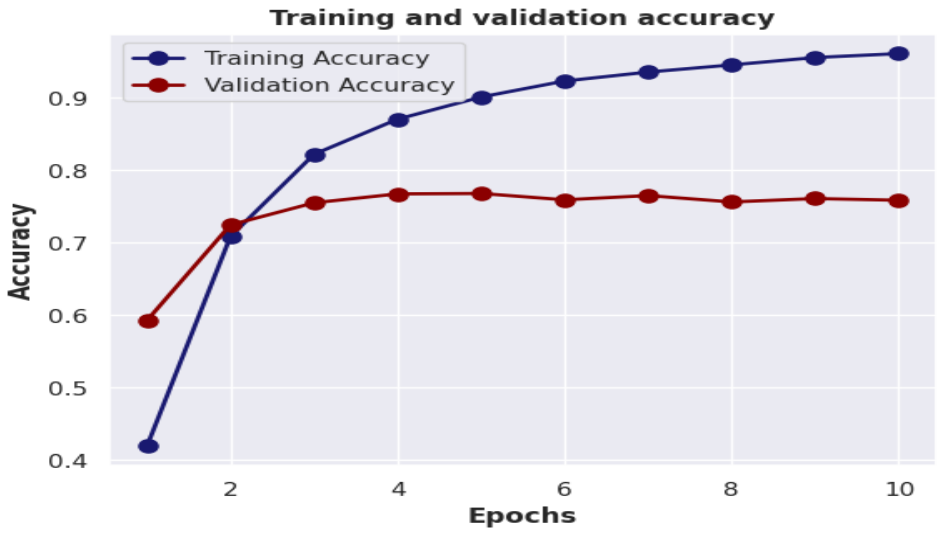

Figure-6: Epochs vs Training & Validation Accuracy Plot for CNN

From the accuracy plot it is observed that, the validation accuracy not improved more than 80%, it is due to multiclass imbalanced classification problem. Moreover, by proper tunning the vocabulary sizes the model performance can be improved.

3.5.2. Validation and Training loss:

Validation loss and training loss were calculated with ten epochs of the model. Validation loss is found to be 1.1975 in epoch-1, validation loss is found to be 0.8874 in epoch-2, 0.8157 in epoch-3, epoch-4 the accuracy is found to be 0.8368, epoch-5 the validation loss is found to be 0.9226, validation loss is found

to be 0.9839 in epoch-6, 1.0309 in epoch-7, validation loss is found to be 1.0566 in epoch-8, validation loss is found to be 1.0979 in epoch-9 and at epoch-10 the validation loss is found to be 1.2932. That is, it is seen here that the validation loss value is upward.

On the other hand, in training loss, the training loss in epoch-1 is 1.7307, in epoch-2 the training loss is 0.9389, in epoch-3 the training loss is 0.5914, in epoch-4 the training loss is 0.4330, training loss in epoch-5 is 0.3245, training loss in epoch-6 is 0.2557, training loss in epoch-7 is 0.2114, in epoch-8 the training loss is 0.1729, in epoch-9 the training loss is 0.1373, training loss in epoch-10 is 0.1214. That is, the value of training loss is downward.

Table-6: Validation and Training loss for CNN

No	Validation Loss	Training Loss
Epoch 1	1.1975	1.7307
Epoch 2	0.8874	0.9389
Epoch 3	0.8157	0.5914
Epoch 4	0.8368	0.4330
Epoch 5	0.9226	0.3245
Epoch 6	0.9839	0.2557
Epoch 7	1.0309	0.2114
Epoch 8	1.0566	0.1729
Epoch 9	1.0979	0.1373
Epoch 10	1.2932	0.1214

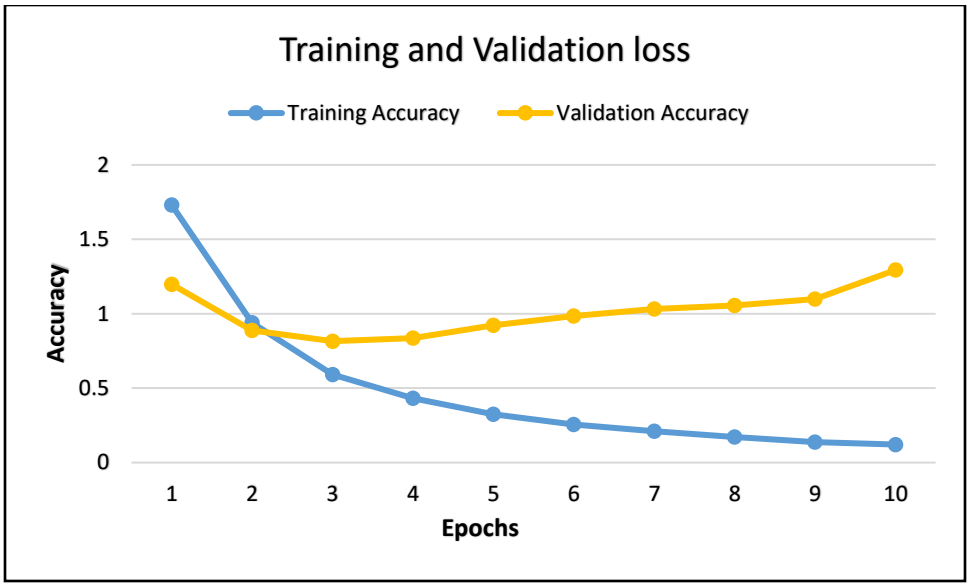

Figure-7: Epochs vs Training & Validation Loss Plot for CNN

3.5.3. Classification Report Model Performance

Table-7: Classification report of CNN Algorithm

	precision	recall	f1-score	support
Accident	82.96	89.60	86.15	125.000000
Art	40.74	51.16	45.36	43.000000
Crime	76.16	79.86	77.97	144.000000
Economics	81.25	53.42	64.46	73.000000
Education	81.32	83.94	82.61	249.000000
Entertainment	79.44	82.13	80.76	207.000000
Environment	56.25	58.44	57.32	77.000000
International	86.32	75.23	80.39	109.000000
Opinion	73.60	57.86	64.79	159.000000
Politics	80.09	86.34	83.10	410.000000

Science	71.88	69.70	70.77	66.000000
Sports	91.06	89.92	90.49	238.000000
Accuracy	78.95	78.95	78.95	0.789474
Macro avg	75.09	73.13	73.68	1900.000000
Weighted avg	79.16	78.95	78.79	1900.000000

By observing precision, recall and f1-score we can see that all the classes are classified reasonably well except Art and Environment.

3.5.4. Confusion matrix

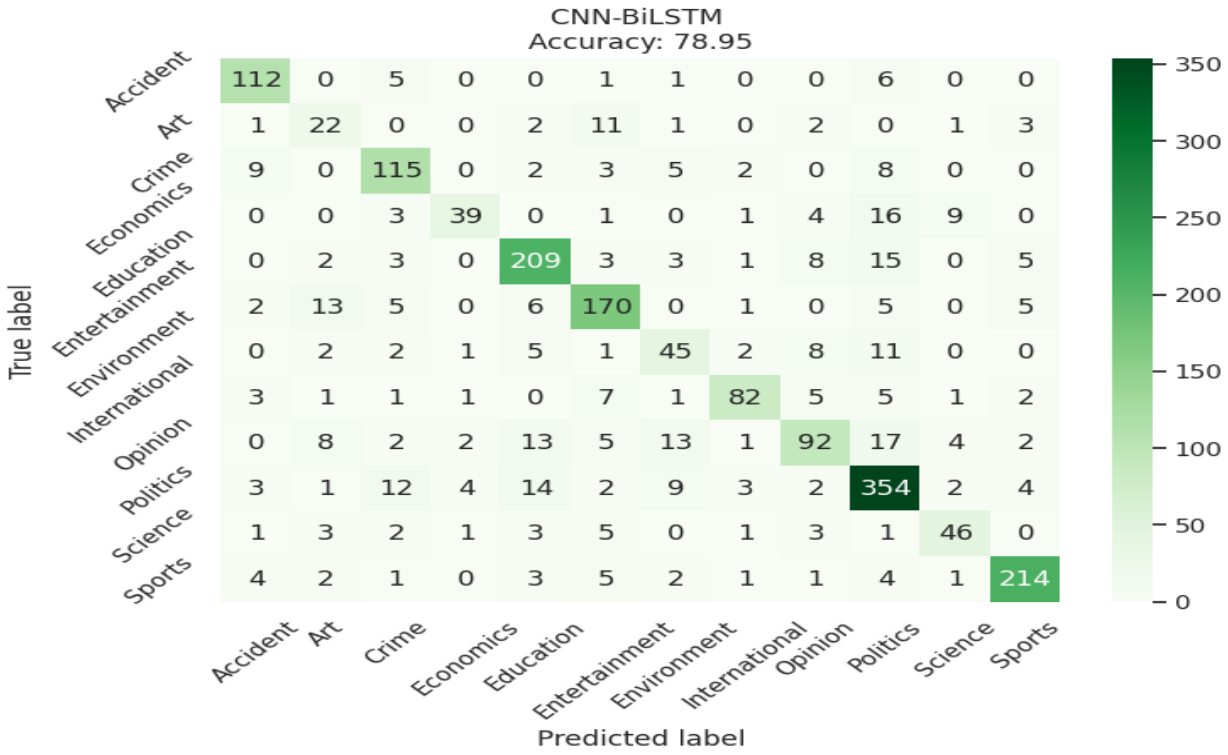

Figure-8: Confusion matrix for CNN

The confusion matrix provides a good understanding about how many documents are correctly classified in each class and which classes get confused during classification. Here, we can see that Art, Entertainment, Politics category gives larger number of false classified result.

3.5.5. Testing CNN algorithm with our own news

Table-8: Test CNN algorithm using dummy news

Sample News	Class	Accuracy %
'নিহত একরামুল হক নয়ন (৩৬) শহরের নর্থ বেঙ্গল এলাকার মৃত আছের আলীর ছেলে। ওই এলাকায় 'আলিফ মেডিকেল স্টোর' নামে তার একটি ওষুধের দোকান আছে। জেলার সহকারী পুলিশ সুপার (সার্কেল) আদিবুল ইসলাম জানান, মঙ্গলবার রাত সাড়ে ১২টার দিকে শহরের এয়ারপোর্ট এলাকায় ছিনতাইকারীরা নয়নকে কুপিয়ে তার মোটরসাইকেল নিয়ে যায়। ঘটনার পর রাতেই পুলিশ অভিযান চালিয়ে পাঁচ ছিনতাইকারীকে আটক এবং নয়নের মোটরসাইকেলটি উদ্ধার করে বলে জানান তিনি। আদিবুল জানান, নয়ন পরিচিত একজনকে মোটরসাইকেলে করে মহেন্দ্রনগর ইউনিয়নের আমবাড়ী গ্রামে পৌঁছে দিয়ে শহরে ফেরার পথে ছিনতাইকারীর কবলে পড়েন। "ছিনতাইকারীরা নয়নকে আটকে তার মাথায় ধারালো অস্ত্র দিয়ে আঘাত করে। পথচারীরা তাকে উদ্ধার করে রংপুর মেডিকেল কলেজ হাসপাতালে পাঠালে সকালে সেখানে তার মৃত্যু হয়।" ছিনতাকারীদলের বাকি সদস্যদের ধরার চেষ্টা চলছে বলে সহকারী পুলিশ সুপার জানান।'	Crime	99.96

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Discussion

In today's world, the amount of news online is increasing day by day. Every day, new news, various types of news are being created on different topics and the amount of news is increasing. The latest news is not only published in newspapers, but also on various online news portals, blogs, social media and various news-based websites. There is a certain measure of newspaper writing that cannot be overwritten. But there is no limit to writing on online websites, blogs, news portals or social media, you can write as much as you want on these online sites. Therefore, besides writing the online news in detail and descriptively, the news must be categorized. As a result, the amount of news has increased tremendously and the amount of news is constantly increasing, besides one thing to be careful is that the news is all classified. If the cumulative news can be classified in some way, then this news can be used in future. We miss a lot of important news because this news is not classified as such. News can be divided into international, national, sports, entertainment, politics, IT, entertainment, education and many more categories. Therefore, if the online news is classified in a digitized manner using machine learning or deep learning, then this news can be used for many important tasks in the future.

In this paper, some algorithms of machine learning and deep learning are applied. Algorithms applied are CNN.

Table-9: Classifier Description

Classifier	Description
CNN	CNN detects unique features from images without any human intervention.

4.2. Experimental Results and Analysis

Table-10: Classifiers accuracy, recall and precision

	Precision %	Recall %	F1_Score %	Support %
Accuracy	78.95	78.95	78.95	0.789474
Macro avg	75.09	73.13	73.68	1900.000000
Weighted avg	79.16	78.95	78.79	1900.000000

Algorithm techniques	Accuracy %	Recall %	Precision %
CNN	78.95	78.95	78.95

CNN algorithm is used in our research paper. Here 78.95% assurance is calculated using CNN algorithm. In addition to predicting news headlines, an experiment has been conducted with a news article to see if news headlines can be predicted correctly. This test shows that they tested with a crime related news and gave correct results. The accuracy of the correct result is 99.96%.

Table-11: Test CNN algorithm using dummy news

Sample News	Class	Accuracy %
'নিহত একরামুল হক নয়ন (৩৬) শহরের নর্থ বেঙ্গল এলাকার মৃত আছের আলীর ছেলে। ওই এলাকায় 'আলিফ মেডিকেল স্টোর' নামে তার একটি ওষুধের দোকান আছে। জেলার সহকারী পুলিশ সুপার (সার্কেল) আদিবুল ইসলাম জানান, মঙ্গলবার রাত সাড়ে ১২টার দিকে শহরের এয়ারপোর্ট এলাকায় ছিনতাইকারীরা নয়নকে কুপিয়ে তার মোটরসাইকেল নিয়ে যায়। ঘটনার পর রাতেই পুলিশ অভিযান চালিয়ে পাঁচ ছিনতাইকারীকে আটক এবং নয়নের মোটরসাইকেলটি উদ্ধার করে বলে জানান তিনি। আদিবুল জানান, নয়ন পরিচিত একজনকে মোটরসাইকেলে করে মহেন্দ্রনগর ইউনিয়নের আমবাড়ী গ্রামে পৌঁছে দিয়ে শহরে ফেরার পথে ছিনতাইকারীর কবলে পড়েন। "ছিনতাইকারীরা নয়নকে আটকে তার মাথায় ধারালো অস্ত্র দিয়ে আঘাত করে। পথচারীরা তাকে উদ্ধার করে রংপুর মেডিকেল কলেজ হাসপাতালে পাঠালে সকালে সেখানে তার মৃত্যু হয়।" ছিনতাকারীদলের বাকি সদস্যদের ধরার চেষ্টা চলছে বলে সহকারী পুলিশ সুপার জানান।'	Crime	99.96

CHAPTER 5
IMPECT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1. Impact on Society

Thanks to the internet, social media has now added a new dimension to modern life. People in rural areas no longer browse through newspapers, they search and read news online through smartphones. Everything that is happening in the country and abroad is available online through YouTube, Messenger, Instagram, LinkedIn, Pinterest, Trampler, Snapchat, Viber, WhatsApp, Facebook, Twitter, Google and other social media. Social media timelines or newsfeeds are getting filled with all kinds of necessary and unnecessary news. 70% of internet users worldwide are connected through social media. Among young people the rate is about 90%. A report shows that about 80% of internet users in Bangladesh use Facebook. The use of internet has increased the level of social communication many times than before. Information, opinions, pictures, videos, etc. of any person are exchanged through various types of data and documents through technology. None of these types of data or documents are classified, so if we search for a specific topic, it does not come up well. So, this is a big disadvantage. Data classification is done to overcome this difficulty. That is, if the news is classified into different categories such as Politics, Education, Sports, Entertainment, Crime, Opinion, Accident, International, Environment, Economics, Science Tech, Art, etc., it will be very beneficial in finding news. As a result, if you search for news by writing any type of keyword, the chances of news of that keyword type will increase many times. So, to read any type of news it is very important to first classify the news into different categories. Classifying it into different categories will make it much easier for people from different walks of life to read news online or through social media.

5.2. Impact on Environment

People have been interested in news since ancient times. People used to read news from newspapers since ancient times. But as the day goes by the advent of technology people don't read news in newspapers anymore. Nowadays all types of news are available through internet and all types of news are available through social media. All kinds of news about what is happening anywhere in the world comes to us in a moment. Any news about the environment through the Meteorological Department comes to us online within a short period of time. Like where in the world there is a cyclone, where there is a flood, where there is a tornado and what is the climate of a region of the world, all kinds of news come to us through online. We can know instantly what the weather is like in any area or any country. We do not look at the

environment or the weather report to decide what work can be done at some point in the future. In this way it is possible to survive from the face of many big accidents or losses. Therefore, the contribution of news in the environment is many.

5.3. Ethical Aspects

Every news media has a specific policy or editorial policy. Media can adopt policies independently. Many times, the media gives wrong information to newspapers online. Eg: It is unethical to provide wrong information online about any person related to political, financial, income tax and other matters. However, people in many places spread negative news against people online. Although today there is less fake news on social media than ever before. That means we all should take care of ethics and use online or social media ethically.

5.4. Sustainability Plan

To put our data set into different classes we need to divide the data or documents into different categories specifically. So that news or documents can be identified in different categories. So that the data can be used in the future and the data can be used for any useful purpose. That's why there should be a long plan, so that any type of data can be pre-processed and divided into different categories by applying different algorithms. That's why it's important to have a plane for long-range thinking.

CHAPTER 6
SUMMARY, CONCLUSION, RECOMENDATION AND IMPLEMENTATION FOR FUTURE RESEARCH

6.1. Summary of the Study

The world is getting modernized day by day. As a result of modernization, something is constantly happening in the world. As something new is invented, we are getting to know it through various means. One of the main reasons for this is the Internet. Today the Internet is the main source of information exchange. All classes of people are now using the internet due to the availability of the internet. As a result, any news spread all over the world in no time. This news is usually available from online news portals through internet. And the online news portals are Google Facebook twitter social media blog WhatsApp Instagram messenger LinkedIn telegram viber printrest etc. A lot of information is available on all these online news portals. But all these types of information or news are generally unusable because they are not classified. As a result, to make all the information or news usable data, the data needs to be classified first. Our data set is classified using deep learning algorithm and divided into 12 categories namely- Politics, Education, Sports, Entertainment, Crime, Opinion, Accidents, International, Environment, Economy, Science-Technology, Industry etc. Data science requires the use of machine learning and deep learning such as CNN LSTM ANN SVM etc. to classify the data. CNN model is used in our paper and 78.95% accuracy is obtained.

6.2. Conclusion

News is not being classified at the same rate as the amount of news on online news portals around the world is increasing. As a result, many usable news are left unusable. If the large amount of news on online news portals can be classified then the unusable data can be made usable and can be used for any need. In our research paper, the data set is divided into three parts namely training data, testing data and validation data. Out of total 18999 cleaning data, 13679 data are taken for training, 1900 data are for testing and 3420 data are taken for validation. The data set is classified using deep learning algorithm and divided into 12 categories namely- Politics, Education, Sports, Entertainment, Crime, Opinion, Accidents, International, Environment, Economy, Science-Technology, Industry etc. 78.95% accuracy was obtained using the CNN model in our paper.

6.3. Implication for Further Study

In this paper we have seen how to classify large amounts of data into different categories. If such a large amount of news can be classified in the manner shown in the paper, then all news can be exploited and used in the future. Different types of algorithms in data science can be classified by applying them in different ways, such as Machine Learning Algorithms (Logistic Regression, Multinational Naive Bayes, Random Forest Classifier, Support Vector Machine), Deep Learning Algorithms (LSTM, BiLSTM, GRU, Uni-Gram, ANN , CNN, RNN) etc In the future, more new algorithms will be used to classify the data This paper shows that only one algorithm namely CNN algorithm is used, but in the future more algorithms will be used to classify such data. And in this paper only news or data is divided into 12 categories but in the future newer categories will be divided. In the future, the accuracy of the algorithm used in this paper will be improved and the news headlines can be accurately predicted.

REFERENCE

1. Singh, Prakash Kumar, and Sanchita Paul. “Deep Learning Approach for Negation Handling in Sentiment Analysis.” *IEEE Access*, vol. 9, 2021, pp. 102579–102592., https://doi.org/10.1109/access.2021.3095412.

2. Rabib, Mohammad, et al. “Different Machine Learning Based Approaches of Baseline and Deep Learning Models for Bengali News Categorization.” *International Journal of Computer Applications*, vol. 176, no. 18, 2020, pp. 10–16., https://doi.org/10.5120/ijca2020920107.

3. Usmani, Shazia, and Jawwad A. Shamsi. “News Headlines Categorization Scheme for Unlabelled Data.” *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, 2020, https://doi.org/10.1109/icetst49965.2020.9080726.

4. Khushbu, Sharun Akter, et al. “Neural Network Based Bengali News Headline Multi Classification System: Selection of Features Describes Comparative Performance.” *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2020, https://doi.org/10.1109/icccnt49239.2020.9225611.

5. Ruichao W., John D., Joe C. May 2014. “Machine Learning Approach To Augmenting News Headline Generation.”

6. Hossain, Syeda Sumbul, et al. “Context-Based News Headlines Analysis: A Comparative Study of Machine Learning and Deep Learning Algorithms.” *Vietnam Journal of Computer Science*, vol. 08, no. 04, 2021, pp. 513–527., https://doi.org/10.1142/s2196888822500014.

7. Santos, António Paulo, et al. “Sentiment Classification of Portuguese News Headlines.” *International Journal of Software Engineering and Its Applications*, vol. 9, no. 9, 2015, pp. 9–18., https://doi.org/10.14257/ijseia.2015.9.9.02.

8. Saha, Uchchhwas, et al. “Sentiment Classification in Bengali News Comments Using a Hybrid Approach with Glove.” *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2022, https://doi.org/10.1109/icoei53556.2022.9777096.

9. Jubaer, A.N.M., et al. “Bangla Toxic Comment Classification (Machine Learning and Deep Learning Approach).” *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, https://doi.org/10.1109/smart46866.2019.9117286.

10. Salehin, Mushfiqus, et al. “Generating Bengali News Headlines: An Attentive Approach with Sequence-to-Sequence Networks.” *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, https://doi.org/10.1109/smart46866.2019.9117554.

11. Barua, Adrita, et al. “Multi-Class Sports News Categorization Using Machine Learning Techniques: Resource Creation and Evaluation.” *Procedia Computer Science*, vol. 193, 2021, pp. 112–121., https://doi.org/10.1016/j.procs.2021.11.002.

12. Islam, Md. Majedul, et al. “Performance Measurement of Multiple Supervised Learning Algorithms for Bengali News Headline Sentiment Classification.” *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, https://doi.org/10.1109/smart46866.2019.9117477.

13. Mohiuddin, Ettilla, and Abdul Matin. “Multilevel Categorization of Bengali News Headlines Using Bidirectional Gated Recurrent Unit.” *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, 2021, https://doi.org/10.1109/acmi53878.2021.9528006.

14. Bhuiyan, Md. Rafiuzzaman, et al. “An Approach for Bengali News Headline Classification Using LSTM.” *Advances in Intelligent Systems and Computing*, 2021, pp. 299–308., https://doi.org/10.1007/978-981-15-9927-9_30.

15. Galal Elsayed, Hoda Ahmed, et al. “A Two-Level Deep Learning Approach for Emotion Recognition in Arabic News Headlines.” *International Journal of Computers and Applications*, vol. 44, no. 7, 2020, pp. 604–613., https://doi.org/10.1080/1206212x.2020.1851501.

16. Tudu, Ronald, et al. “Performance Analysis of Supervised Machine Learning Approaches for Bengali Text Categorization.” *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, 2018, https://doi.org/10.1109/apwconcse.2018.00043.

17. Jahara, Fatima, et al. “Automatic Categorization of News Articles and Headlines Using Multi-Layer Perceptron.” *Intelligent Computing & Optimization*, 2022, pp. 155–166., https://doi.org/10.1007/978-3-030-93247-3_16.

18. Bogery, Raghad, et al. “Automatic Semantic Categorization of News Headlines Using Ensemble Machine Learning: A Comparative Study.” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 11, 2019, https://doi.org/10.14569/ijacsa.2019.0101190.

19. Ahmed Foysal, Md. Ferdouse, et al. “Bengali News Classification Using Long Short-Term Memory.” *Advances in Intelligent Systems and Computing*, 2021, pp. 329–338., https://doi.org/10.1007/978-981-33-4367-2_32.

20. Ke Yahan, R. Qu, Lu Xiaoxia. January 2022. "Classification Of Fake News Headline Based On Neural Networks".

Document Viewer

Turnitin Originality Report

Processed on: 30-Dec-2022 10:00 +06
ID: 1987396545
Word Count: 8569
Submitted: 1

Bengali Documents categorization Using Deep L... By Mst. Eshita Khatun

Similarity Index	Similarity by Source	
7%	Internet Sources:	3%
	Publications:	6%
	Student Papers:	2%

exclude quoted | exclude bibliography | exclude small matches | mode: quickview (classic) report | print | refresh | download

1% match ("Intelligent Computing & Optimization", Springer Science and Business Media LLC, 2022)
"Intelligent Computing & Optimization", Springer Science and Business Media LLC, 2022

1% match ("Emerging Technologies in Data Mining and Information Security", Springer Science and Business Media LLC, 2021)
"Emerging Technologies in Data Mining and Information Security", Springer Science and Business Media LLC, 2021

1% match (student papers from 30-May-2022)
Submitted to Exeed College on 2022-05-30

1% match (Internet from 29-Dec-2021)
https://www.researchgate.net/publication/228084521_Text_Classification_Using_Machine_Learning_Techniques

<1% match (Internet from 11-Sep-2022)
https://academic-accelerator.com/Manuscript-Generator/Decision-Tree/Random-Forest-Classifier

<1% match (Internet from 11-Jun-2022)
https://academic-accelerator.com/Manuscript-Generator/Naive-Bayes/artificial-neural-network

<1% match (Internet from 11-Jun-2022)
https://academic-accelerator.com/Manuscript-Generator/Naive-Bayes/support-vector-machine

<1% match (Syeda Sumbul Hossain, Yeasir Arafat, Md. Ekram Hossain. "Context-based News Headlines Analysis: A Comparative Study of Machine Learning and Deep Learning Algorithms", Vietnam Journal of Computer Science, 2021)
Syeda Sumbul Hossain, Yeasir Arafat, Md. Ekram Hossain. "Context-based News Headlines Analysis: A Comparative Study of Machine Learning and Deep Learning Algorithms", Vietnam Journal of Computer Science, 2021

<1% match (Ronald Tudu, Shaibal Saha, Prasun Nandy Pritam, Rajesh Palit. "Performance Analysis of Supervised Machine Learning Approaches for Bengali Text Categorization", 2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2018)
Ronald Tudu, Shaibal Saha, Prasun Nandy Pritam, Rajesh Palit. "Performance Analysis of Supervised Machine Learning Approaches for Bengali Text Categorization", 2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2018

<1% match (student papers from 01-May-2022)
Submitted to Australian National University on 2022-05-01

<1% match ("Machine Learning for Predictive Analysis", Springer Science and Business Media LLC, 2021)
"Machine Learning for Predictive Analysis", Springer Science and Business Media LLC, 2021

<1% match (Internet from 28-Aug-2022)
https://link.springer.com/article/10.1007/s40031-021-00630-5?code=1e65fbf1-2b35-4554-8038-840e1e64acaa&error=cookies_not_supported

<1% match (Internet from 20-Jan-2022)
https://link.springer.com/article/10.1007/s00170-020-05620-3?code=55bd048e-04ae-4942-8fb1-7d01972c0c55&error=cookies_not_supported

<1% match ("Machine and Deep Learning in Oncology, Medical Physics and Radiology", Springer Science and Business Media LLC, 2022)
"Machine and Deep Learning in Oncology, Medical Physics and Radiology", Springer Science and Business Media LLC, 2022

<1% match (Xi Yang, Hui Li, Rong Chen. "Underwater image enhancement with image colorfulness measure", Signal Processing: Image Communication, 2021)
Xi Yang, Hui Li, Rong Chen. "Underwater image enhancement with image colorfulness measure", Signal Processing: Image Communication, 2021

<1% match (Internet from 24-Nov-2022)