

**USING DEEP LEARNING TECHNIQUES TO IDENTIFY FALSE NEWS
IN BANGLA**

BY

MD SAKIB KHAN

ID: 161-15-7321

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

Supervised By

MR. MAHIMUL ISLAM NADIM

Lecturer

Department of Computer Science and Engineering
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

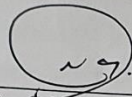
DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

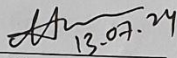
This Project titled “**Using Deep Learning Techniques To Identify False News In Bangla**”, submitted by Md Sakib Khan and to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13 July 2024.

BOARD OF EXAMINERS



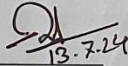
Dr. S.M Aminul Haque (SMAH)
Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Board Chairman



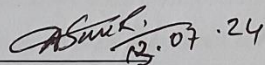
Ms. Nazmun Nessa Moon (NNM)
Associate Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner 1



Mr. Deawan Rakin Ahamed Remal (DRAR)
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner 2



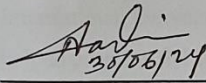
Dr. Abu Sayed Md. Mostafizur Rahaman (ASMR)
Professor
Department of CSE
Jahangirnagar University

External Examiner

DECLARATION

We hereby declare that this research has been done by us under the supervision of **Mr. Mahimul Islam Nadim**, Lecturer, Department of Computer Science and Engineering Faculty of Science and Information Technology, Daffodil International University. We also declare that neither this research nor any part of this research has been submitted elsewhere for the award of any degree.

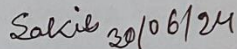
Supervised by:



30/06/24

Mr. Mahimul Islam Nadim
Lecturer
Department of CSE
Daffodil International University

Submitted by:



30/06/24

Md. Sakib Khan
ID: 161-15-7321
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

At First we would like to express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year thesis successfully.

We really grateful and wish our profound our indebtedness to **Supervisor Mr. Mahimul Islam Nadim**, Lecturer, Department of CSE, Faculty of Science and Information Technology, Daffodil International University, Dhaka. His Deep Knowledge & keen interest with supportive instructions helped us in the field of machine learning based research, finally we completed our work on “**Using Deep Learning Techniques To Identify False News In Bangla**”. Their 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.

We would like to express our heartiest gratitude to **Mr. Mahimul Islam Nadim**, Lecturer and **Dr. Sheak Rashed Haider Noori** Professor and Head, Department of Computer Science and Engineering, Faculty of Science and Information Technology, DIU, for his valuable support and advice to finish our project and also heartiest thanks to other faculty member and the staff of department of CSE, Daffodil International University.

At last, again we want to thank all the good wishers, friends, family, seniors for all the help and inspirations. This research is a result of hard work and all those inspirations and assistance.

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

ABSTRACT

The prevalence of fake news has increased the need for sophisticated detection algorithms, particularly those that operate in languages other than English. The purpose of the study is to address the issues surrounding Bangla, an indigenous tongue that is viewed as less important. It is advised to use a full data with approximately 50,000 articles to do this. Many deep learning models have been tested with this dataset, including hybrid architectures, reversible gated recurrent units (GRU), long short-term memory (LSTM), and 1D convolutional neural network models (CNNs). Several useful criteria, including retention, accuracy, F1 score, and efficiency, were used in this study to assess the model's efficacy. A sizable software was used to do this. By using the Bidirectional GRU models, we are able to achieve an amazing 99.14% accuracy rate in identifying erroneous information in Bangla. We carry out comprehensive tests to prove the effectiveness of these models in this regard. The significance of preserving dataset balance and the amount of continuous improvement work needed are highlighted by our research. By employing a prototype online application and a web application machine learning model classifier to identify fake news in Bangla, this study establishes the foundation for future developments in the detection process. In particular, it makes a major and low-resource contribution to the advancement of Bangla false news detection systems.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
CHAPTER 1: INTRODUCTION	1-6
1.1 Introduction	1
1.2 Motivation	2
1.3 Rational of the study	3
1.4 Research Questions	3
1.5 Expected Output	4
1.6 Project Management and Finance	4
1.7 Report Organization	6
CHAPTER 2: BACKGROUND	7-11
2.1 Terminologies	7
2.2 Related Work	7
2.3 Research Summery	10
2.4 Challenges	10
CHAPTER 3: RESEARCH METHODOLOY	12-27
3.1 Introduction	12
3.2 Research Subject and Instrumentation	12
3.3 Proposed Methodology	13

3.4 Data Collection Procedure	15
3.5 Statistical Analysis	17
3.6 Implementation Requirements	26
CHAPTER 4: EXPERIMENTAL RESULT AND DISCUSSION	28-35
4.1 Introduction	28
4.2 Experimental setup	28
4.3 Experimental result & Analysis	28
4.4 Discussions	34
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	36-39
5.1 Social Impact	36
5.2 Social Environment	36
5.3 Ethical Aspects	37
5.4 Sustainability plan	38
CHAPTER 6: CONCLUSION AND FUTURE RESEARCH	40-41
6.1 Summary of the study	40
6.2 Conclusion	40
6.3 Possible Impacts	41
6.4 Implication for Further Study	41
REFERENCES	42-44

LIST OF FIGURES

FIGURES	PAGE
Figure 3.1 About Recommended Process.	14
Figure 3.2 True (1.0) sample data.	16
Figure 3.3 False (0.0) sample data.	16
Figure 3.4 Word Visualization Training Data.	18
Figure 3.5 Word Visualization Testing Data.	18
Figure 3.6 Bangla Stop vocabulary	20
Figure 3.7 Example of Tokenization for Bangla words.	21
Figure 3.8 Data about class distribution for training and testing.	21
Figure 3.9 BGRU approach general architecture	22
Figure 3.10 The general structure of the 1D CNN model.	23
Figure 3.11 The general structure of the LSTM model.	23
Figure 3.12 The general architecture of the CNN + GRU hybrid model.	24
Figure 3.13 The general architecture of the CNN + LSTM + Global Max Poling 1D hybrid model.	24
Figure 3.14 Proposed Architecture of the Bidirectional GRU model.	25
Figure 4.1 CM and the ROC-AUC Curve for bidirectional GRU.	31
Figure 4.2 CM and the ROC-AUC Curve for 1D CNN.	31
Figure 4.3 CM and the ROC-AUC Curve for LSTM.	32
Figure 4.4 CM and the ROC-AUC Curve for Hybrid CNN + GRU.	32
Figure 4.5 CM and the ROC-AUC Curve for Hybrid CNN + LSTM + GlobalMaxPooling1D	33

FIGURES	PAGE
Figure 4.6 Detecting bogus news in Bangla with a web prototype.	33
Figure 4.7 Real news detection utilizing web prototype.	34
Figure 4.8 Fake news detection utilizing web prototype.	34

LIST OF TABLES

TABLES	PAGE
Table 3.1 Dataset Column Description	16
Table 4.1 Before Balancing Dataset.	29
Table 4.2 After Balancing Dataset.	30

CHAPTER 1

Introduction

1.1 Introduction

In this age of digital media, news media play a critical role in shaping public opinion and choosing products. The rapid spread of information via online forums is problematic, though, because of fake news. Disseminating untrue data may lead to actual aberration, the spread of false data, and even civil unrest when lies or incomplete facts are presented as truth. Determining false information and opposing it is one of the main challenges to the dissemination of information [1]. Many research has focused only on cultural dominants, abandoning minority languages and communities with little resources and solutions, despite the significant effort made to counter false news. Of the three mentioned languages, Bangla is one that millions of people speak fluently [2]. Nonetheless, the complexity of Bangla presents unique difficulties when it comes to spotting misleading material. The unique terrain formed by language variation, colloquial expressions, and cultural sensitivities may provide challenges for traditional approaches [3]. Our goal was to close this gap by examining fake news in Bangladesh through the application of deep learning techniques. The goal is to optimize the performance of a neural network known as a gate recurring unit (GRU) in conjunction with other deep learning techniques [4]. In a variety of natural language processing (NLP) applications, such as sentiment analysis and translating languages, this approach has shown promising results.

Our study is predicated on a set of articles that were specifically created to evaluate the veracity and accuracy of news from Bangladesh. This dataset offers a strong foundation for additional research and analysis and is a crucial tool for comprehending the complexity of false news in Bangladesh. We are able to capture news breakdowns from various online distribution channels thanks to a number of factors in the dataset. Our rigorous methodology makes the findings of our studies discoverable and accessible. First, we use Word Clouds to represent the most often occurring terms after analyzing the raw data. Finally, we prepare the data, which includes tasks like term removal and correction. This ensures that the data fed into our model is meaningful, unambiguous, and noise-free.

Creating and utilizing deep learning algorithms is our research's primary goal. In our research, we employed a variety of models, such as numerous hybrid deep learning models, 1D convolutional neural network models, and long short-term memory systems. After conducting a comprehensive

research and analysis, we found a paradigm that, when applied to Bangladesh, showed the highest reliability and accuracy in false cases being identified. In addition, we talk about the need of addressing socioeconomic inequality as a common issue in the detection of fake news. Regression analysis is used to take these differences into account and assess how they affect the efficacy of the model. We are carrying out research and trying to determine the best strategies for lowering inaccurate results and false negatives—two critical measures of the accuracy of false detection methods.

Our research is an extensive endeavor to use annotating databases and DL to increase the quality of fake news in Bangladesh. By investigating the nuances of detecting fake news in this country and maximizing the effectiveness with the two-track GRU model, we want to contribute to the ongoing discussion on combating deceptive data on the web. In the sections that come next, we bridge the gap among Bangladesh's particular challenges with false news identification and modern technology by explaining our path, discoveries, and ideas.

Our efforts have contributed the following:

- Analyze the efficacy of the bidirectional Gated Recurrent unit (GRU) in determining the accuracy of Bangla news stories.
- Addressing this problem of class disparity using a variety of modeling techniques and improving the accuracy of the model in identifying authentic from false information.
- Bridging the gap between in-depth, sophisticated study and the unique challenges presented by the Bengali language, providing practical solutions to important issues.

1.2 Motivation

The modern technology age frequently faces problems with misleading and deception. A healthy society, political concord, and public opinion might all suffer as a result of them. DL, ML, and AI encourage us to strive toward preserving the knowledge base within the community at large. To help everyone steer clear of false information, we need to identify and combat fake news. After giving the task some thought, we were unable to come up with whatever that could satisfy the inquiry's standards. In light of this, we sought advice from one of our esteemed instructors.

The system of institutions we had chosen for our study's issue was authorized by him, and he suggested that we proceed forward in this direction as it is imperative that misleading information

be prevented in Bangladesh. Consequently, we have decided to focus on the topic " **Using Deep Learning Techniques To Identify False News In Bangla.**" We also note that in today's environment, fake news usually takes advantage of local and regional peculiarities. The detecting procedure takes into account the distinctive features of Bengali culture because of Bangla-specific technology. This made us want to do further investigation on the subject. Close relationships exist between the global environments in which both the ml and dl approach's function.

1.3 Rationale of the study

The classification of items and the detection of fake news have been the subjects of extensive investigation. The topic of " **Using Deep Learning Techniques To Identify False News In Bangla.**" has few completed publications. Thus, we employ the web application Flask in our study to detect patterns using many algorithms and classifications. To keep the project going forward, we have developed a specialized model and classification for the subject as a consequence of our meticulous planning. Ensuring that our Bengali-speaking population receives accurate and trustworthy data is essential for maintaining public confidence in the media and public discourse. As a teaching tool, developing a false news detecting gadget fosters critical thinking and media comprehension among Bengali-speaking consumers. The populace that speaks Bangla has the confidence to participate in virtual conversations because there is a device available to verify the accuracy of content written in Bengali. Furthermore, because of DL classifiers and methodologies, fraudulent news in Bengali may be identified with great accuracy using Web Flask and machine learning techniques.

1.4 Research Questions

This study has been completed with a great lot of enthusiasm and effort. Completing this project proved to be difficult for us as well. Developing a solution that is accurate, workable, and equitable faces a number of challenges. In order to make sense of these concepts and provide a solution, the researchers want to ask the following questions:

- Is it possible to access unprocessed image data to support research?
- Is it possible for me to preprocess the data using a machine approach?
- Can these tactics reduce the dissemination of misleading information?

- In what ways might the results of this study and the previously described techniques help people?

1.5 Expected Output

Included in this part are a few facts that are assumed to represent our lowest-case expected outcome. The real conclusion of this Bangla fake news was predicted using ml and dl utilizing a number of classification algorithms based on the findings of the research that follows. The objective of this research project is to construct a rigorous, efficient process or procedure that can discriminate between real and fake news produced by an online application. This process will be based on a hypothesis that was created through training then dataset assessment. This leads to an inventory of the anticipated outcomes as follows:

- We'll demonstrate whether or not the info is true after looking at its content.
- A better understanding of how to use DL and ML to identify fake news.
- By utilizing label data and out-of-date content, we want to compare our findings with those of other studies.
- Based on all of the information presently available, figuring out the most effective technique for spotting fake news from Bengali using ml & dl.

1.6 Project Management and Finance

Effective project management and sound financial oversight are essential for any enterprise to succeed and endure over time. That particularly applies to the project that is being presented, "Using Deep Learning Techniques To Identify False News In Bangla." All of the careful planning, arranging, and carrying out of tasks—from data gathering to model development and evaluation—are included in project management. A comprehensive project schedule that identifies significant checkpoints, prudently distributes resources, and sets deadlines will be developed in order to promote a systematic and effective workflow. The feasibility of the project will also be significantly impacted by careful fiscal oversight, which includes planned spending on resources for computers, equipment, and possible research partnerships. Using resources as efficiently as possible while upholding fiscal accountability can be achieved by implementing a careful and transparent financial strategy. With the aid of these unified approaches for estimating finance and leadership, the researchers are going to be able to effectively negotiate the complicated waters of

the study, make sure that the assets designated are in accordance with the intended results of the study, and enable the project to move smoothly toward its objectives.

1.7 Report Organization

The first chapter served as an illustration of how to make the endeavor more understandable by stressing its perseverance, drive, challenges with the study, and anticipated results. The report goes into additional depth about the document's general structure in this section.

Chapter 2 discusses the progress achieved in this area thus far. The final segment of the subsequent section exemplifies the thoroughness resulting from its focus on this subject. The main challenges or limitations of the investigation are briefly discussed. The next part contains sections on connected works, a summary of the subject, and challenges that required to be solved in order to finish the assignment.

It provides a philosophical assessment of the research endeavor in Chapter 3. To address the psychological aspect of this study, more details on the statistical techniques used are provided in the next paragraph. The sequential steps involved in DL approaches for identifying fake news utilizing algorithms based on machine learning will be illustrated in this part through the usage of a concept for a web application. This chapter provides a detailed overview of the dataset acquisition method and the tools employed for data preparation. Furthermore, in order to examine the prediction model, the subdivision's ultimate sacrifice provides a complex matrix assessment together with an accuracy markup for the classification method. More study has to be done in order to be sure that machine learning (ML) has been used to detect and classify false news in Bangladesh and that methods based on deep learning have been used to assess actual accuracy. This section covers the research subject and tools, methodology, data collection and analysis tactics, suggested model, training techniques, and practical demands that must be fulfilled for the undertaking to be completed. Each of the study's strategies includes a thorough rationale that covers all dl approaches and ml categories.

Chapter 4 includes a performance assessment, an examination of the study findings, and the study's results themselves. To help bring the project to life, this part provides some labeled data sets and

experimental material. This is followed by an analysis of the formulation and outcomes of both the ml and dl approaches.

The research overview, project follow-up, and conclusion were included in Chapters 5 and 6. This section's goal is to clarify why every project report item complies with the specification. The surroundings, society, and sustainability are affected. We discussed the areas in which we failed and how they would impact people in related fields in the future as we concluded this segment.

CHAPTER 2

Background

2.1 Terminologies

Often, the introduction provides a summary of the study, duties, or scientific difficulties. In "Associated Work," we will look at other writers' question papers to determine whether their approaches and accuracy match ours. I'll discuss the approaches, veracity, and substance of further research articles that are pertinent to this inquiry in an article on similar studies. I'll give a brief overview of our relevant research in the study section. For the part on problems, I have chosen to examine every problem that we managed to address during the course of this research, in addition to how we improved the layer's quality and precision. Every problem has already been examined in terms of using ML and DL algorithms to classify label material as authentic or fraudulent.

2.2 Related Works

To aid in the automated detection of bogus news, Zobaer Hossain et al. [5] offered a dataset comprising 50,000 Bangla news articles. They looked at neural network and traditional methods, highlighting SVM's (91% F1-score) better performance than neural models like CNN or Bi LSTM. The majority of the collection is made up of Bangla news articles. As per the findings of Risul Rasel et al. [6], we blended newly acquired fake news data with pre-existing datasets to generate a Bangla fake news collection. The research employed a range of models for machine learning (ML) (LR, SVM, KNN, and MNB, the Ada increase, DT), deep machine learning (LSTM, Bi LSTM, CNN), transducer (Bangla-BERT, m-BERT), as well as deep neural networks (LSTM, Bi LSTM, CNN). Curiously, CNN turned out to be the most successful organization, with reliability rates of 95.7 % and improving recall for data containing inaccurate information. Utilizing feature extraction techniques like Vectorizer with Count and TF-IDF Vectorizer, 4678 distinct news items comprised the dataset. Accuracy of the models rose between 1.4% to 3.4% when tested against datasets that were made accessible to the public. Techniques to enhance the detection of fake news in Bangla were provided by Md Hossain et al. [7] by accounting for imbalanced datasets. Using a skewed dataset containing 97% majority instances, Ban Fake News, they applied oversampling techniques such as Randomized Excessive collection, SMOTE, and ADASYN. The researchers got a 79.1% the F1 score without stacking refinement and a 93.1% F1-score using SMOTE.

Notably, without these tactics, the starting point models would give an F1-score of 67.6%. Although the article focuses only on Bangla, it is a significant step forward in the identification of bogus news.

Md Ali et al. [8] proposed a more effective machine learning strategy to detect bogus news in Bengali. By combining categories for fake news and legitimate news in one dataset, they hoped to boost the recognition rates. Their Python-based solution produced an amazing 98.08% efficiency using a range of methods for feature extraction and machine learning algorithms. Nonetheless, the study emphasizes the necessity for more investigation into the feasibility of spotting fake news in Bengali on social media. Liza Fahmida et al. present a multilingual fake news detection tool that can concurrently identify fraudulent information in Bengali and English [9]. They use TF-IDF and n-grams Analyses to extract features while comparing the outcomes of six automated learning techniques.

With an accuracy of 93.29% and an F1 score of 0.93, the model performs admirably when utilizing its Linear Support Vector Classify approach. Even though the dataset includes items written in both Bengali and English, the study's primary contribution is a revolutionary strategy that can be used to media outlets in both languages. The dataset isn't easily accessible, though, which is one disadvantage. A transfer learning-based approach that leverages BERT's masked LM technique is presented by Sourav Saha et al. [10] for the purpose of identifying fake news in Bengali. Pre-trained BERT is used to obtain an accuracy of 95% and specificity rate, surpassing methods including as CNN, SVM, NB, and LSTM.

The study's methodology involves the application of SVM, MNB, masked LM (MLM) preliminary processing, and ReLU activation. It's interesting that the article only addresses the identification of bogus news in Bengali. Manal Iftikhar, Arshad Ali, et al. [11] offer a machine learning method for automatically identifying bogus news. They gather information from several internet sources and categorize it using SVM, naive Bayes, and LSTM approaches in order to distinguish between legitimate and fraudulent news. Their method produces a choice tree classifier based on the ISOT false news data that has 99% accuracy, which is an amazing result. Moreover, ensemble techniques like Support Vector Machine (SVM) and Random Forest (RF) are used. In particular, the accuracy of the suggested SVM + RF model is 2.3% higher than that of the current model.

The authors look into English fake news related to COVID-19 using ensemble techniques. They attain a remarkable 98% accuracy by combining BERT with Roberta. Leela et al. [12] provide an

example of a false news detection system that employs the Logistic Regression (LR) strategy and compares its dependability to the Support Vector Machine (SVM) technique. On the 311-examples of data that will be used in the study, LR performs better than SVM (91.68%) in determining the presence of bogus news, with a precision of 95.12%. The identification process used combined LR and SVM algorithms, with LR turning out to be more accurate. However, the study does not adequately address the system's flexibility to handle other datasets or domains, nor the dataset's source.

An innovative technique for categorizing Bengali news using models from machine learning and deep learning is presented by Tamim Al Mahmud & Antara Mondal [13]. Their process include collecting over 20,000 organized data points, executing numerous designs, and selecting the best performing one. Several metrics are used in the technique to assess the performance of the model, such as F-measure, accuracy, precision, recall, and recall. Interestingly, they achieve a success rate of 87.12% using the SVM (bag of words) technique. Although the research provides an accuracy-enhancing strategy, one limitation of its previous work is the small model scope. Among the models used are gated recurrent units, long short-term memory, stochastic forests, support vector machines, and linear support the vector classifiers.

In this study, Elias Hossain et al. [14] investigate the use of deep learning and machine learning in the identification of fake news in Bangla. Employing an archive of 57,000 Bangla news articles, they examine techniques such as Gated Recurrent Unit (GRU) and Bi-LSTM with Wrapped & Fast Text. The study highlights the superiority of deep learning over conventional machine learning. The Bi-LSTM model achieves a robust 96% accuracy, while the GRU model only achieves 77%. The report also includes a comparison of earlier studies and experimental findings. The proposed method shows adaptability for real-time news classification. However, the study's relevance to other countries is limited to the identification of bogus news in Bangla. Iftikhar Ahmad, Suhail Yousaf, Muhammad Ovais, and Muhammad Yousaf [15] share an automated learning ensembles strategy for automatically classifying news articles as real or fraudulent. Using ensemble methods and multinational feature sets across many domains, the research explores discrete textual properties. Utilizing datasets from Kaggle and these ISOT Fake News Collection, the authors evaluate ensemble and linear SVM approaches. The results showed that the ensemble learner approach is superior than individual students, with the bagged classifications (decision trees) performing best with 94% accuracy. The recall performance of the boost classifier (XG

Boost) is extremely good (95.25%), while the recall performance of the bagging classifier is very good (0.942). XG Boost's efficient error detection process is the source of its superior performance. A limitation of the research is that it assesses algorithms from several fields using a single dataset.

2.3 Research summary

I spent a great deal of time researching the many options the community offers. To evaluate the reliability of the information set, five distinct methods are integrated. Furthermore, employing a ml and dl framework for a web application, a Passively Aggressive Classifier is employed to forecast the label of content in order to categorize and identify fake news in Bengaluru. I applied many techniques to my distinct dataset. In this instance, a dataset I put together from the internet served as the main source of statistics. Using the exact same source, I am going to be able to evaluate the validity of the six techniques I used and investigate the effects of the additional data I supplied. The source material which was used to combine the previous data has been improved to become a perfect duplicate of the first. Tags can be classified under related categories and labels, which clarifies their significance. With Python acting as the primary engine of choice, our extracting features and data preparation methods used methods such as DL and Passive Aggression Foundation approaches for classified data to anticipate fake news.

2.4 Challenges

Integrating the processing of all of these ensembles appeared to be the primary challenge of this project, given how challenging it was to control these four enormous data sets. A variety of instruments and techniques were utilized to purify and harmonize the dataset. We were left waiting a while to see the finished result because of the magnitude of the raw data and the simple fact that they included several layers from different eras. There are more data sets available in this field. Because we didn't complete the study, I have to spend a lot of effort into every activity linked to it, which makes it difficult for me to swiftly coming up with the best answers.

Applying DL and artificial intelligence (AI) models to categorize the content of the Bangla false news into an internet-based prototype presented additional challenges in managing this massive number of data as well as in the planning for categorization. The development of natural language processing (NLP) algorithms to automatically categorize occurrences as false news has garnered significant attention due to the possible advantages of identifying and mitigating such activity.

However, this endeavor is not without its difficulties. Since they harm and undermine their targets, their convoluted and phony news may also make it challenging for an NLP systems to distinguish between fake and real.

- Gathering data;
- Preparing data.
- Achieving 90% accuracy at the greatest degree feasible.
- Making a decision about each DL model based on test results.

CHAPTER 3

Research Methodology

3.1 Introduction

In this study area, we want to discuss the full process and the methods used to classify the detection of false news in Bangla. The collection and analysis of data, as well as the proposed model—which is further supported by the relevant calculation, graph, surface, and description—are some of the fundamental components. The best accuracy DL classification models are used in this study's data preprocessing and prediction utilizing internet datasets. Rounding off the chapter are a summary of the statistical ideas we employed in our research and a set of simple implementation needs. In this study, we built our models using four different kinds of classes. Generally speaking, this Bangla news collection has a wide variety of material types. For every kind of directory in the databases, there are two label classes that we concentrated on in our work: true and false. For each file group, those participating in our research were told to utilize a maximum of two labels.

3.2 Research subject and equipment

The topic of a research project is typically one that is currently being investigated and assessed by others, meaning that it has potential for model creation, data collection, value determination, and model dissemination. We discuss the instruments and techniques we use for measurement. We coded in Python and used a number of additional libraries, like NumPy, SkLearn, OpenCV, etc. I conducted all of my research and monitoring using the Co Lab search engine interface. Python is used by Google's Colab for programming, and statistical techniques corresponding to ML and DL are applied to categorize labels like true or false utilizing web applications.

Libraries used:

- **Matplotlib:** Among Matplotlib's visualization features is the the Pyplot program graphical graphing suite of methods. This might be used, for example, to define the boundaries of an outline while generating shapes, or to show corners in a tale.
- **NumPy:** Vector manipulation became popularized via the NumPy program for Python. Linear math, including vector calculations, the Fourier coefficient evolve, and the indexing transformation, which is are included in the topic material. The the NumPy module when

Python provides a wealth of tools and methods for handling arrays of different formats. Devices may be designed logically and realistically with NumPy. The NumPy is a module for Python made specifically for quantitative evaluation, to put it simply. Another name for this statement is "computation. Py."

- **Scikit-learn:** This effective and user-friendly program is used to assess anticipated data. Three Python tools—NumPy, SciPy, and Matplotlib—that may be downloaded for free and customized to meet the demands of each user were used in its creation.
- **Seaborn:** The thing that comes next Matplotlib is used by the well-known Python visualization application Seaborn. This is a simple-to-use tool for illustrating statistics in an artistic way.
- **H5py:** The h5py module of the Python software gives users the ability to use HDF5 binary data format. NumPy is used to manage and store large quantities of data, especially integers, in HDF5.
- **TensorFlow:** TensorFlow seems to be a free machine learning toolkit. It provides an extensive range of tools for developing and putting into practice various machine learning simulations, including NN. TensorFlow's versatility, efficiency, and usability make it an excellent choice for a wide range of artificial intelligence applications.
- **Pandas:** Pandas provides a freeware toolkit for analyzing information and manipulation for a certain language. It provides basic data formats and statistics techniques to facilitate the management of data that is organized, including summary information.
- **OS:** Developers can engage with the computing system that these individuals use thanks to the many features of the case of Python OS module.

3.3 Proposed Methodology

The evidence reviewed in this study could have been chosen using a variety of approaches or working strategies. Choosing a model, gathering and processing data, creating word clouds, eliminating stop words, sanitizing the text, and carrying out a performance evaluation employing the internet-based application as the result were all steps in the workflow for this project.

Phase 1: Data Collection: I got the data from websites since not every kind of fact has its own repository at this time. Recently, a subsequent list of data classifications was created.

Phase 2: Preparing the Data: Every piece of information was gathered in an independent manner, utilizing every available method. There are many examples of inappropriate and halting phrases there. Before developing it, I literally handle the last phase of the chosen dataset.

Phase 3: Get the dataset ready: Data extension and preprocessing are still carried out according to label and content. For training reasons, we need to suspend word removal, stop word visualization, and tokenize it. We only completed the most essential and least time-consuming steps to get the data suitable for splitting.

Phase 4: Choose a model: We select a prediction technique, train that, and then assess it employing my data in order to boost reliability. DL uses many distinct filters. Only one model was ultimately selected to evaluate the accuracy of the data after several models were used to improve machine design performance and for the classification of inaccurate reports using ML model.

Phase 5: Performance Assessment: This section addresses all ramifications. Following training, testing, and validation, these techniques gave us a limited degree of reliability for the two class labels in the data. Apart from the confusion matrices, accuracy data were also generated and displayed. To find fake news in Bangla, utilize Web Flask and machine learning models.

Phase 6: Conclusion and Further Findings: A continuous history and a summary will be included with this.



Fig 3.1: About Recommended Process.

The procedures in our study approach are displayed in Fig. 3.1 and can help Bangladesh identify misleading information. We used a variety of data relating to the media. We collected data from an openly accessible dataset made up of Bangladeshi news articles. After carefully reviewing this data gathering, we cleaned up the phrasing and removed any unnecessary terms to ensure that this data set would only include accurate and relevant data. Word clouds are employed to make frequently occurring words easier to understand. Using pre-existing data, we build and improve models to explore the subject of deep learning. Among these models are mixtures of long short-term memory (LSTM) [18], unidirectional gates for recurrent units (GRU) [16], and 1D convolutional neural networks (CNN) [17]. After a careful analysis, we determined which models were most useful for identifying erroneous data in Bangla language details. Furthermore, we use permutation strategies to handle the issue of class diversity in order to ensure fair representation and improve overall model performance. In addition to giving an overview, our approach searches for ways to reduce the number of positive results that are false and inaccurate outcomes. We are able to combat deceptive data in the Bangladeshi linguistic context by employing a comprehensive method that blends linguistic expertise with deep learning technologies.

3.4 Data Collection

The dataset used in this study was sourced from the freely available Kaggle [19] repository. Each of the four primary files in our suggested database provides clarification on fake news in Bangla.

1. Real-48K.csv and Illegitimate-1K.csv: The article ID, news source website, publishing date, group, title, news content, and label (one for true, zero for false) are crucial fields in these files. The bulk of our data is included in these files, which contain a variety of authentic and fake news articles.

2. LabeledAuthentic-7K.csv and LabeledFake-1K.csv: The news pieces have greater context and comprehension thanks to these files. Columns such as F-type (a sort of fake news found only in LabeledFake-1K.csv), relationship (associated or unconnected), and reference (the organization that might verify the news claim) make up these. These files enhance our dataset by including elements necessary for a comprehensive analysis.

The dataset consists of 10 columns labeled "True and False" and 58,478 content data in total was constructed. The dataset's train and test sections are divided apart. Test has 8,501 content and label information points, compared to 49,977 in Train.

	articleID	domain	date	category	source	relation	headline	content	label
0	1	bd-pratidin.com	2018-09-20 08:16:43	Sports	আফগান ক্রিকেট বোর্ড (এসিবি) প্রধান	Related	হঠাৎ আফগান ক্রিকেট বোর্ড প্রধানের পদত্যাগ	ক্রিকেট বিশ্বের নতুন চমকের নাম আফগানিস্তান। কয়...	1.0
1	2	jugantor.com	2018-09-20 20:20:20	Sports	Reporter	Related	টস হেরে বোলিংয়ে বাংলাদেশ	এশিয়া কাপের ষষ্ঠ ম্যাচে বাংলাদেশ দলের বিপক্ষে ...	1.0
2	3	bd24live.com	2018-09-20 16:39:40	National	Reporter	Related	রাজধানীতে বিশেষ অভিযানে আটক ৪৩	রাজধানীতে মাদক বিরোধী বিশেষ অভিযান পরিচালনা কর...	1.0
3	4	bd24live.com	2018-09-19 18:27:56	National	জননিরাপত্তা বিভাগের সচিব	Related	উক্কানি রোধে নজরদারি থাকবে সামাজিক যোগাযোগ মাধ্য...	সনাতন ধর্মাবলম্বীদের সবচেয়ে বড় ধর্মীয় উৎসব দুর...	1.0
4	5	somoynews.tv	2018-09-20 10:15:28	Finance	পেট্রোল পাম্প মালিক সমিতি সভাপতি	Related	'যেখানে তেল আসত ৭ দিনে, এখন তা আসবে অতি দ্রুত'	ভারত থেকে পাইপ লাইনের মাধ্যমে সরাসরি দেশে জ্বা...	1.0

Fig 3.2: True (1.0) sample data.

	articleID	domain	date	category	headline	content	label
0	1	channelhdhaka.news	2019-03-14T13:34:14+00:00	International	মুরগির হামলায় শেয়াল নিহত	বাংলায় একটা প্রবাদ আছে, শেয়ালের কাছে মুরগী বর্...	0
1	2	earki.com	সেপ্ট ১৭, ২০১৮	Miscellaneous	বিটিভিতে যেবার আমি ইন্টারভিউ দিতে গেলাম	BTV থেকে লোকজন আসছে, ইন্টারভিউ নিবে। চারজনের টি...	0
2	3	earki.com	২০:৩৯, জানুয়ারি ১৪, ২০১৯	Miscellaneous	বিদেশ থেকে উন্নতমানের বিরোধীদল আমদানি করার পরা...	অদ্বুত বিরোধীদলহীনতায় ভুগছে সরকার। এ এক অন্যরক...	0
3	4	channelhdhaka.news	2018-06-30T15:56:47+00:00	Sports	অবসর নেয়ার ঘোষণা দিলেন মেন্সি !	রাশিয়া বিশ্বকাপ নকআউট পর্বে ফ্রান্সের সাথে ৪-৩...	0
4	5	motikonho.wordpress.com	2013-03-05T21:55:45+00:00	Miscellaneous	মাদারফাকার নহে, ব্রাদারফাকার: সাকা দৈনিক মতি...	নিজস্ব মতিবেদক মাদারফাকার নহে, আমি ব্রাদারফাকার...	0

Fig 3.3: Fake (0.0) sample data.

Below is an overview of Table 3.1, containing a list of all the columns in these files:

Table 3.1: Dataset Column Description

Column Name	Description of the Column
Article ID	News ID
domain	news publisher sites name
date	Date of Publication
category	News category
source	News sources
relation	Related or Unrelated

headline	News headline
content	News article
label	1 as authentic or real, 0 as fake
F-type	Fake news type (only in LabeledFake-1K.csv)

Encoder for labels:

Whenever they have one or more columns, datasets containing multiple characteristics are commonly used by ML and DL professionals. Words and numbers may be combined to create these unique IDs. In an effort to render the data easier to comprehend or accessible to consumers, words are frequently used to categorize training data.

Encoding the labels is one method of converting signs into text that can be read by computers. To create a structure which reflects numbers, identifiers must be transformed as part of this previously outlined process. Whether or not to apply such labels is ultimately up to the developers who develop ML and DL systems. For development under monitoring at this point, pre-processing of the organized data is necessary.

3.5 Statistical Analytics

3.5.1 Analyzing Data

The goal of our data pre-processing is to transform the unprocessed textual data into an arrangement suitable for evaluation and model training. The necessary "content" and "label" fields must first be extracted using real and fake news databases. The collected samples are then concatenated to create a collection of datasets that can be used for testing and training. After assembling our information, we began with typographical adjustments, such as removing superfluous letters and punctuation. These activities aim to improve the appropriateness of word choice within the text. Word Cloud representation is an animated depiction which renders the structure of language easier to understand by emphasizing the bulk of word repetitions. Tokenization is a technique that turns words into vectors of numbers that are easily combined by deep learning models [20]. Our ability to conduct in-depth research is enhanced by this meticulous approach to data processing, which also makes it easier for our algorithms to identify incorrect information in Bangla.

3.5.2 Word Cloud Visualization

Words with great visual impact are important for our datasets for training and testing as well as for our statistical analysis, as the figures. 3.4 and 3.5 show. This technique provides a visual representation of the major phrases used in the report's content, which makes it easier to identify patterns and trends in the vocabulary used [21]. Significant insights on implicit and distinguishing characteristics can be gained. To help guide the process of producing our data, the visual is a helpful tool. The subsequent material presents a realistic picture of explicit growth in language in Bangla while also making it feasible. With this simple technique, we can connect the dots among raw data and meaningful analysis, which aids in our decision-making process while creating an a false-positive research strategy.



Fig 3.4: Word Visualization Training Data.



Fig 3.5: Word Visualization Testing Data.

3.5.3 Stop word Removal

We recognize the language used in our efforts to improve the quality of the content. Even though they are commonly used, the phrases stated above usually have no linguistic context and have no significance inside the classification system. To overcome this issue, we use stop phrases, which we received from GitHub [22] and use a custom library called "Stop words Bengali (BN)". The library provides a wide range of Bengali symbols that had been meticulously preserved to align with the grammar of our language. By integrating these libraries, we can eliminate superfluous characters and enhance our data processing efficiency. As the following examples show, we eliminate punctuation often in order to make our text classification efforts more logical:

ওরা, কখনও, ছিলেন, জন, তোমার, থাকবে etc.

```
# Loading bengali stop words file
with open('/content/drive/MyDrive/Fake bangla/stopwords-bn.txt', 'r', encoding='utf-8-sig') as f:
    stop_word = f.read()
    stop_word = stop_word.replace(" ", "")
    stop_word = stop_word.split('\n')
    print(stop_word)
```

['অতএব', 'অথচ', 'অথবা', 'অনুযায়ী', 'অনেক', 'অনেকে', 'অনেকেই', 'অন্তত', 'অন্য', 'অবধি', 'অবশ্য', 'অর্থাৎ', 'আই', 'আগামী', 'আগে',

Fig. 3.6. Bangla Stop vocabulary.

3.5.4 Text Cleaning

We carry out text correction, an essential stage in the dataset augmentation process. This strategy comprises two main methods designed to raise the standard and relevance of written content. News is first sent via a filter system that only retains articles that are longer than a predetermined word count of 100. Our commitment to delivering informative and compliant high-quality material is ensured by filtering methods. Moreover, a text correction technique is applied to systematically modify the content by removing unnecessary characters such as tabs, breakpoints, special markings, etc. The text has been edited to remove some Bengali characters, line breaks, and typical symbols. By use of an extensive preservation process, we ensure that the information we gather will be ready for close inspection. Thanks to this improvement, our software can now quickly and reliably assess the reliability of Bangla news.

3.5.5 Tokenization and Padding

Tokenization & padding are essential elements of modern data use. Tokenization, or converting written material into numerical sequences, provides our model with the capacity to understand language [23]. The connection between written language and mathematical symbols is created by assigning a unique integer value to each word. Padding improves the chance of convergence during patterns training by ensuring that each sequence has a constant length [24].By converting utterances to mathematical models and confirming that sequence length varies so frequently, our algorithms are able to analyze the complex linguistic nuances of text. In order for our computers to properly learn from the data and distinguish between real and fake Bangla news, this phase is essential.



Fig 3.7: Example of Tokenization for Bangla words.

3.5.6 Data Preparation

We weren't able to randomly split the data for model training and testing throughout the data prep phase. The most significant "content" and "label" columns—a total of 49,000 items—are extracted from the real and fake news databases and used as a training dataset. The label_authentic and label_fake csv files were also concatenated to provide a testing dataset of 8,000 articles overall.

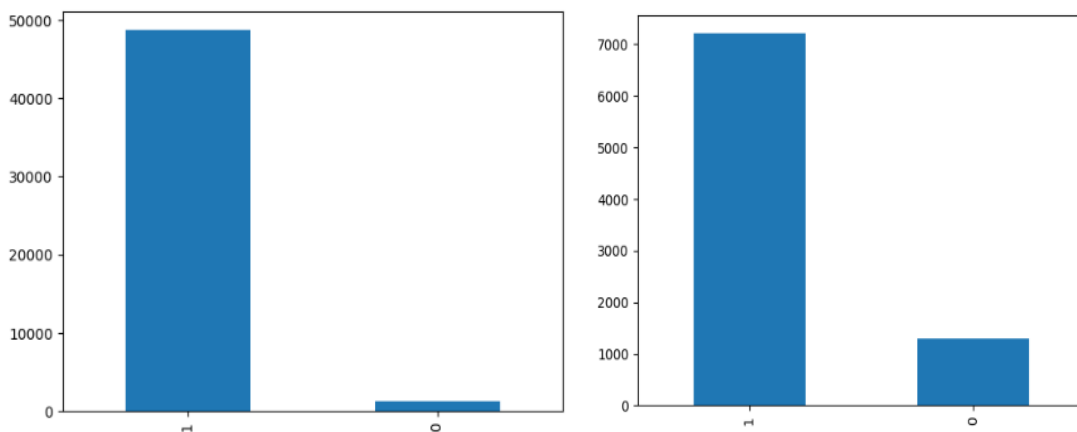


Fig 3.8: Data about class distribution for training and testing.

3.5.5 Applied Models

In order to comprehend Bangla news trustworthiness, our study employs many methods. We analyze the following deep learning models: CNN + GRU, CNN + LSTM + Global Max Poling 1D, Bidirectional Gated Recurrent Units (GRU), and Long Short-Term Memory (LSTM), CNN, and CNN + GRU. We use many models to capitalize on their ability to capture language variations,

patterns, and context. This extensive analysis helps us achieve our goal of creating a reliable system for identifying bogus news in Bangla.

1. **Bidirectional Gated Recurrent Unit:** Initially, we use an unidirectional GRU system to successfully record subsequent patterns by combining data from both sides. Three layers make up the model: global average pooling, bidirectional GRU, and sticking. ReLU's dense activation layers prevent dropouts excessive fitting and permit nonlinear modifications. The output layer uses sigmoid activation to provide binary outcomes for classification.

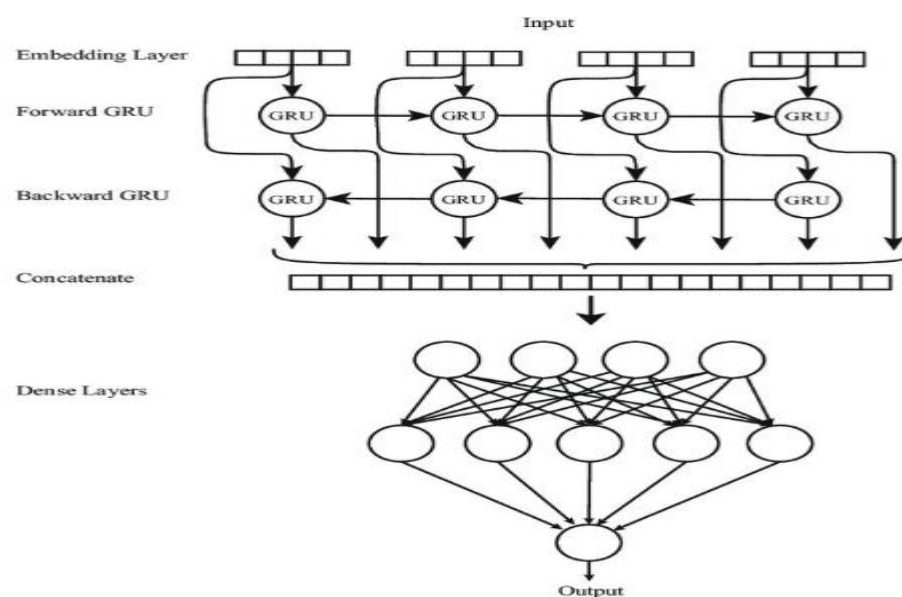


Fig 3.9: BGRU approach general architecture.

2. **1D Convolutional Neural Network:** 1D CNN models are also included in our analysis. It comprises of four layers: smoothing out embedding, MaxPooling, and 1D convolutional for separating features. Deep layers and later on dropout strategies prevent overfitting and improve the model's generalizability.

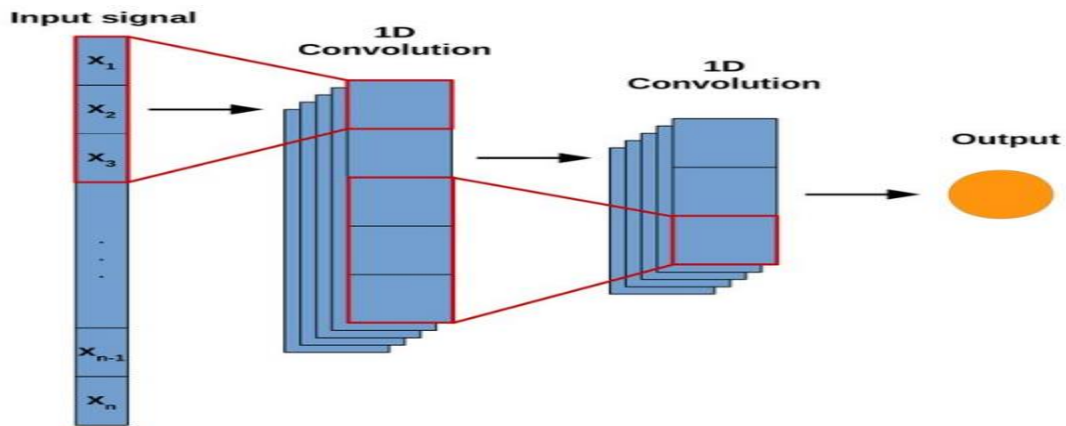


Fig 3.10: The general structure of the 1D CNN model.

3. **Long Short-Term Memory:** We also address 1D CNN models in our analysis. Smoothing, 1D the convolutional, maximal pooling, and embedding layers are used for feature extraction. The model is made more generalizable and protected against overfitting by deep layers and following dropout strategies.

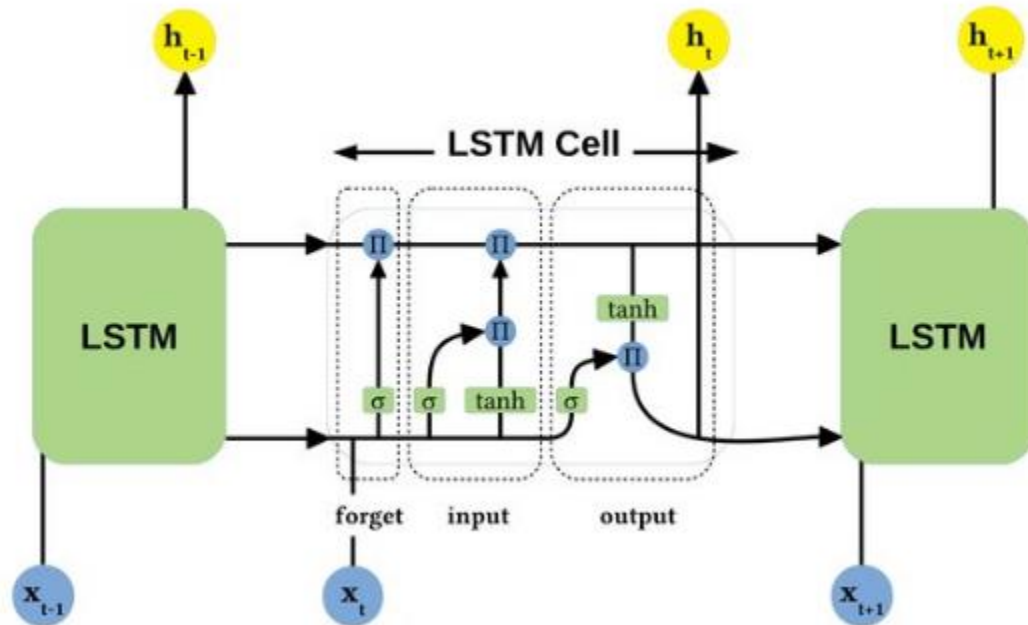


Fig 3.11: The general structure of the LSTM model.

4. **CNN + GRU:** The aforementioned model uses a hybrid approach as well and blends convolutional layers with gated recurrent units (GRU). GRU layers require sequential connections; features are eliminated in convoluted layers. The tiny fragments of the two materials are linked before reaching the Dense layers.

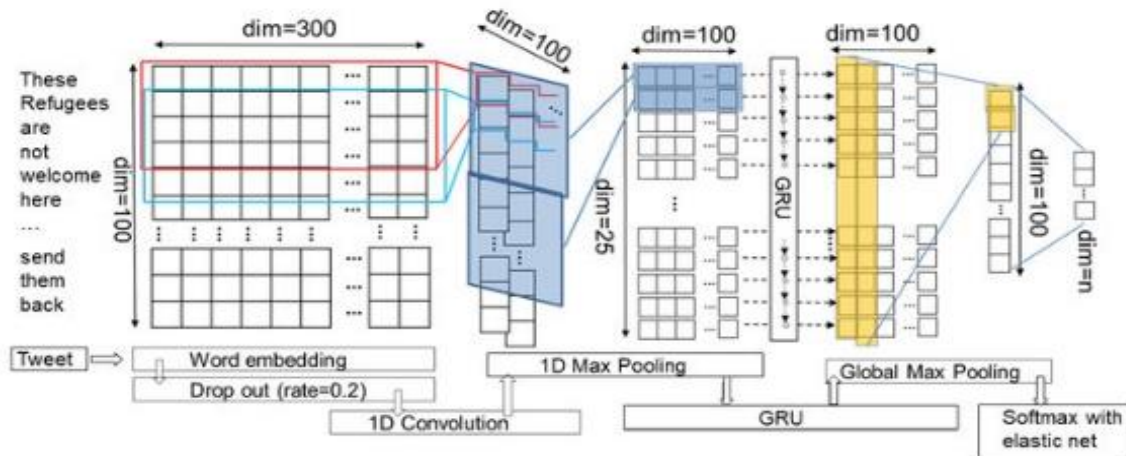


Fig 3.12: The general architecture of the CNN + GRU hybrid model.

5. **CNN + LSTM + Global Max Poling 1D:** We continuously generate new concepts through the joint modeling of LSTM, GlobalMaxPoling1D, and The convolutional layers. The sequencing assessment of LSTM, the parameter extraction properties of convolutional layers, and the global reference capturing of GlobalMaxPooling1D are all included in this hybrid structure.

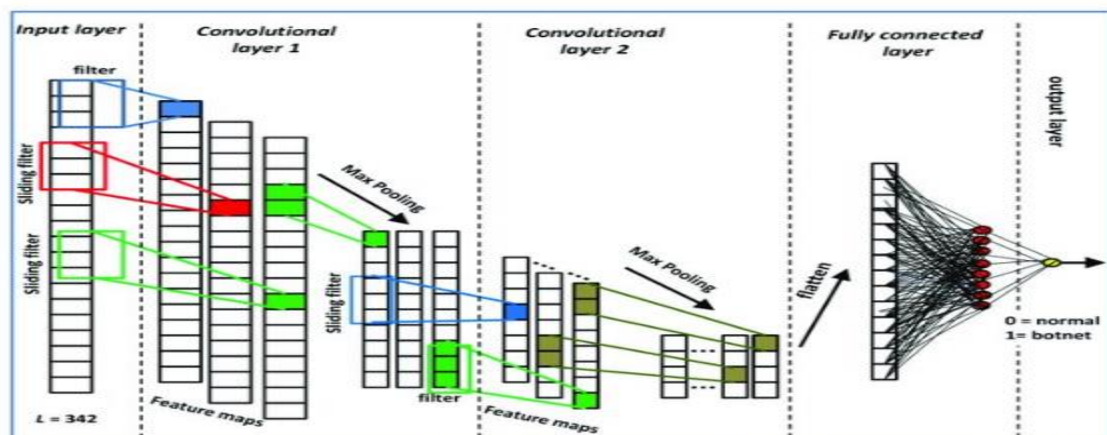


Fig 3.13: The general structure of the CNN + LSTM + Global Max Poling 1D hybrid model.

6. **Passive-Aggressive Classifier:** Binary classification issues may be tackled with the use of the passive aggressive extraction e-learning technique. Because it can educate itself throughout a full set of data while adjusting to new points of data one at a time, it is classified as a "on the internet" algorithm. This flexible method may come in handy when teaching and classification are required in real-time or very quickly, and data is supplied in batches. When quick updates to fresh data are required, this is very common.

3.5.6 Proposed Model

Our proposed design uses a bidirectional gated recursive unit (GRU) to estimate the accuracy of Bangla news items. This approach allows the model to take use of GRU's bidirectionality to effectively gather context in both directions. Figure 3.14 provides an excellent illustration of the architecture.

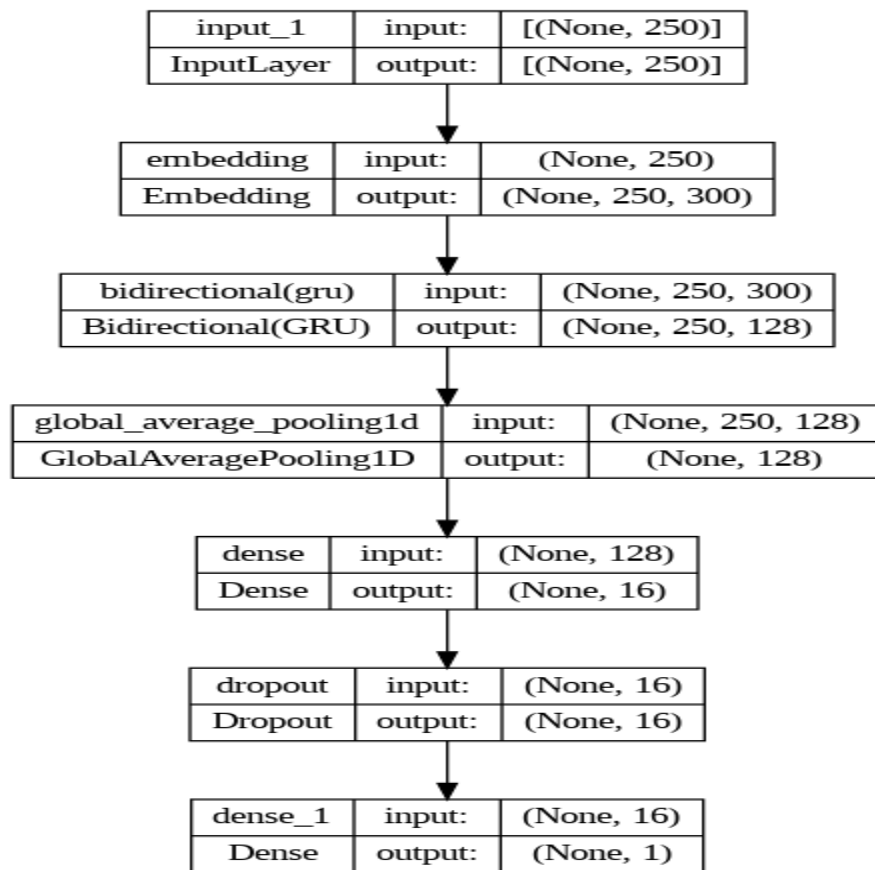


Fig 3.14: Proposed Architecture of the Bidirectional GRU model.

First, text input is converted into a mathematical model via an embedding layer, which fuels the model's first phases. Memory holding capacity is provided by the bidirectional GRU layer, a kind of recursive layer, which is located in the middle of the design. By keeping track of the terms that occur prior to and following the model, this layer aids in understanding its context. The GlobalAveragePooling1D layer stores context; additional dense ones that come after add nonlinearity. By reducing overfitting, rates of dropout improve generalization. Information may be binary classified as true or false because to the use of sigmoid activation in the final output layer. This example acts as the foundation for our attempt to provide a trustworthy false detection method for the Bangla language.

3.6 Implementation Requirements

To verify accuracy, we must use the data collection after completing the earlier steps. Our task was split up into 10 parts: setting up keys. To achieve our goal, each of those actions needs to be completed.

- Gathering data.
- Pre-processing of data.
- Word clouds are visualized.
- Avoid removing words.
- Cleaning text.
- Preparing the data.
- Padding and tokenization.
- Algorithm application.
- Talk about the result and accuracy.
- Implement the Web Flask applications to identify whether a piece of news is "labeling" as fake or real.

Prior to putting the approach into practice, we need to work on its programming. Overall, we examined the accuracy of five different methods. We assessed the procedure's efficacy once it was completed. We looked at the precision and chose the one that was most likely to be useful for our objective. A web-based prototype that uses ML classifiers to identify false information has been eagerly awaited. The prerequisites that must be fulfilled are getting ready. essential

for this sort of job, which entails swiftly classifying false information in Bangla after a precise examination of all pertinent theoretical or numerical concepts and techniques. Potential results that could be required include:

1. Equipment and Software specifications

- Operating System (Windows 7 or above)
- Hard Disk (minimum 1 TB)
- Ram (Minimum 4 GB)

2.Implementation Tools

1. Python Environment
2. PyCharm
3. Google Colab
4. VS code

CHAPTER 4

Experiment Results and Discussion

4.1 Introduction

The approach used to classify bogus news in Bangla is covered in this section. The entire process of developing a model consisted of only a few steps: selecting a model, obtaining and analyzing data, presenting word clouds, removing the words, cleaning the text, and assessing performance based on web application results. We provide the experiment outcomes, which are discussed in the next paragraph.

4.2 Experimental setup

To perform thorough investigations, the hardware, software, and data resources used in this study's setup—" **Using Deep Learning Techniques To Identify False News In Bangla.**"—have been thoughtfully configured. High-performance computers are computing systems with a computational core made up of powerful GPUs and a powerful CPU. I know that no innovation can lead to flawless outcomes. Similarly, to improve accuracy during training, we might change the parameters of our model. By using multiple techniques, we find, however, that the dependability level is fairly high. This provides as an illustrated summary of the assignments that we are currently doing. Recall, accuracy, F1 score, supports, heat maps, and other statistics are displayed in these visualizations. Prior versions of Python modules, arrays, and material classification methods were employed. Notable aspects of the data collection include the DL assessment of a dataset made up of four folders containing contents which may be anticipated for a future set of data according to two labels, such as True and False. The present investigation again employed machine learning (ML) to select the proper category label for informational content and employed an Apache Python web application to detect fake news in Bangla.

4.3 Experiment result and Analysis

The technique we employed determined the various outcomes we obtained. We used five deep learning systems to effectively anticipate the bangla false news label. We used these approaches to calculate each powerplant's efficacy inside the framework, and then we used several verification procedures to offer the forecast's conclusion. Each model utilized a single file once the final dataset was chosen, which included information from our own experience as well as publicly accessible

information gleaned from internet data sources. After completing the dataset strategy, we used Py and its pre-built libraries to test the algorithms' performance. The next step is to determine if the content would be classified as genuine or fake news in Bangladesh using a comparable dataset utilizing Web Flask.

Here, researchers thoroughly analyzed a number of models using significant performance metrics. The metrics include accuracy, precision, recall, and F1 score and give a comprehensive view of the models' performance.

Table 4.1: Before Balancing Dataset.

Models	Accuracy	Precision	Recall	F1 Score
Bidirectional GRU	99.09%	99.13%	99.94%	99.53%
1D CNN	98.94%	99.00%	99.91%	99.45%
LSTM	98.80%	99.03%	99.73%	99.38%
CNN + GRU	98.62%	99.11%	99.47%	99.29%
CNN + LSTM + GlobalMaxPooling1D	98.75%	99.09%	99.63%	99.36%

We test our model and present the results in Table 4.1 above before balancing the dataset. The bidirectional GRU model's remarkable accuracy score of 99.09% is worth mentioning. With an accuracy rating of 98.94%, the 1D CNN model is deemed sufficient. The corresponding accuracy of the LSTM model is 98.80%. Other hybrid models, including CNN + GRU and CNN + LSTM + GlobalMaxPooling1D, also do well in some parameters, like accuracy (98.62%) and score (98.75%). The total of these findings demonstrates how well Bangladeshi false news is identified by our suggested algorithm.

Table 4.2: After Balancing Dataset.

Models	Accuracy	Precision	Recall	F1 Score
Bidirectional GRU	99.14%	99.27%	99.85%	99.56%
1D CNN	99.02%	99.22%	99.77%	99.49%
LSTM	98.68%	99.18%	99.46%	99.32%
CNN + GRU	98.43%	99.21%	99.17%	99.19%
CNN + LSTM + GlobalMaxPooling1D	98.63%	99.24%	99.35%	99.29%

The results of the model assessments using a well-balanced set of data are shown in Table 4.2. Measurements like accuracy scores may be used to evaluate how well the model performs in a balanced setting. The two-phase GRU simulation has 99.14% accuracy. Including recall and F1 score values, the accuracy of the 1D CNN model is 99.02%. The LSTM model has an accuracy rating of 98.68%. These metrics show that the CNN + GRU & CNN + LSTM + GlobalMaxPooling1D algorithms perform well, with 98.43% and 98.63% accuracy scores, correspondingly. The findings show that our models perform better when applied to a balanced dataset, which greatly improves their ability to detect false information in Bangladesh.

After the data are analyzed, we balanced each set of data to show the overall performance of our models, showing the confusion matrix (CM) [29] and ROC-AUC curve [30] for every model. Figures 4.1–4.6 display these diagrams, which illustrate how models are categorized as true positive, genuine negatives, false favorable, and false negative. Naturally. The ROC-AUC turns, which also display the trade-off among the rate of true positives and the proportion of false

positives, illustrate the model's ability to distinguish between valid and false positives. The following charts provide a comprehensive overview of the ways in which models react with fake news in Bengali in an adaptive and efficient manner.

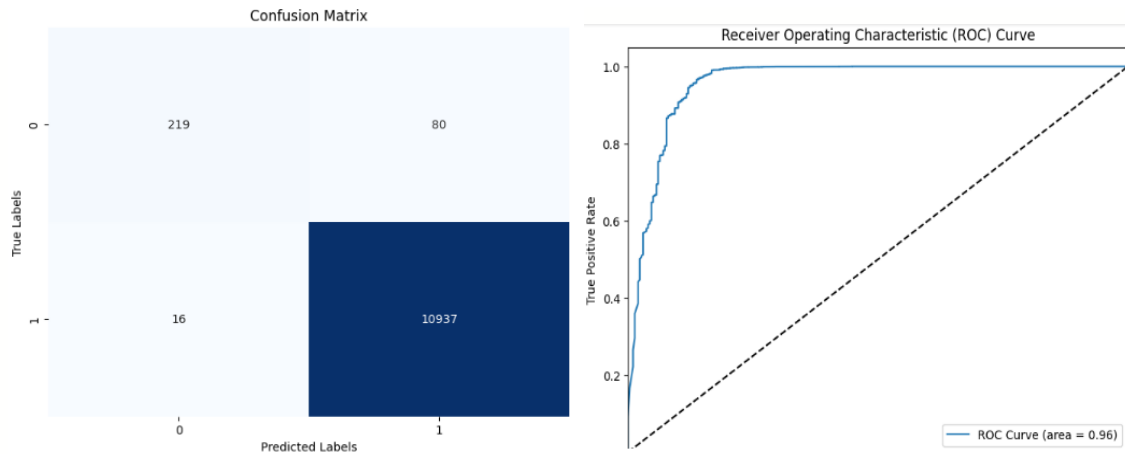


Fig 4.1: CM and the ROC-AUC Curve for bidirectional GRU.

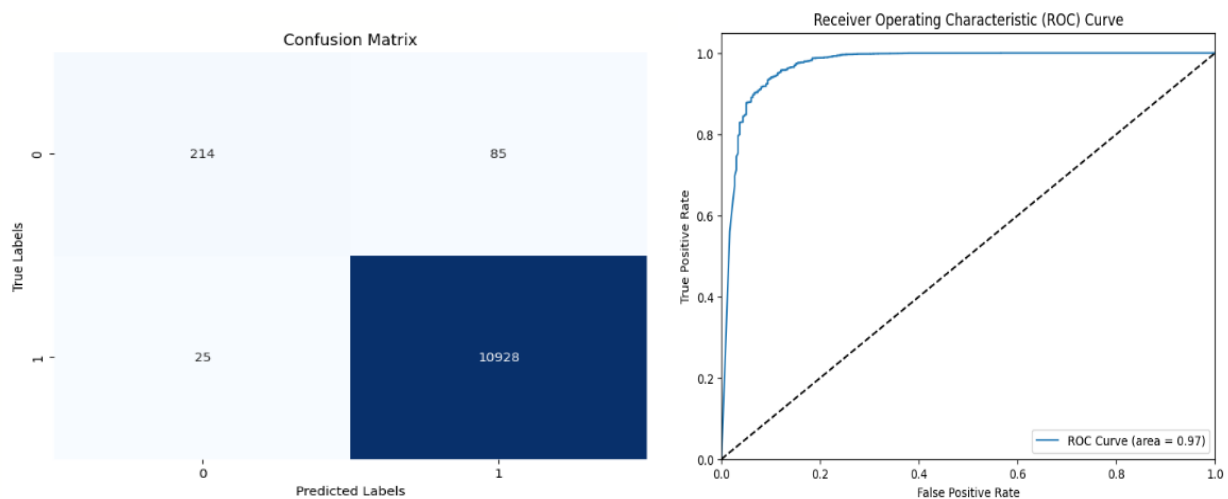


Fig 4.2: CM and the ROC-AUC Curve for 1D CNN.

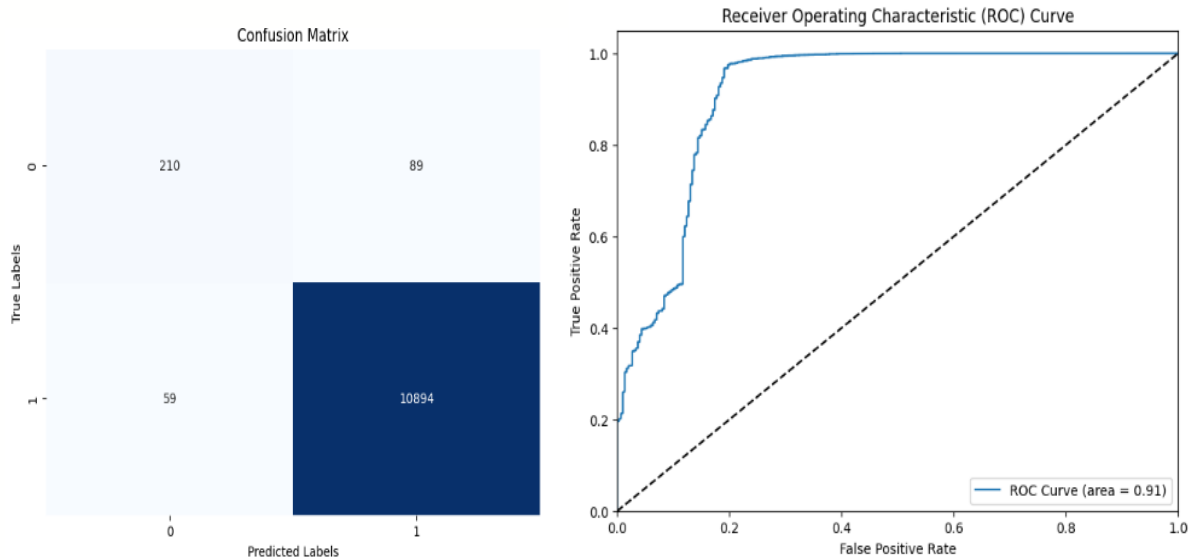


Fig 4.3: CM and the ROC-AUC Curve for of LSTM.

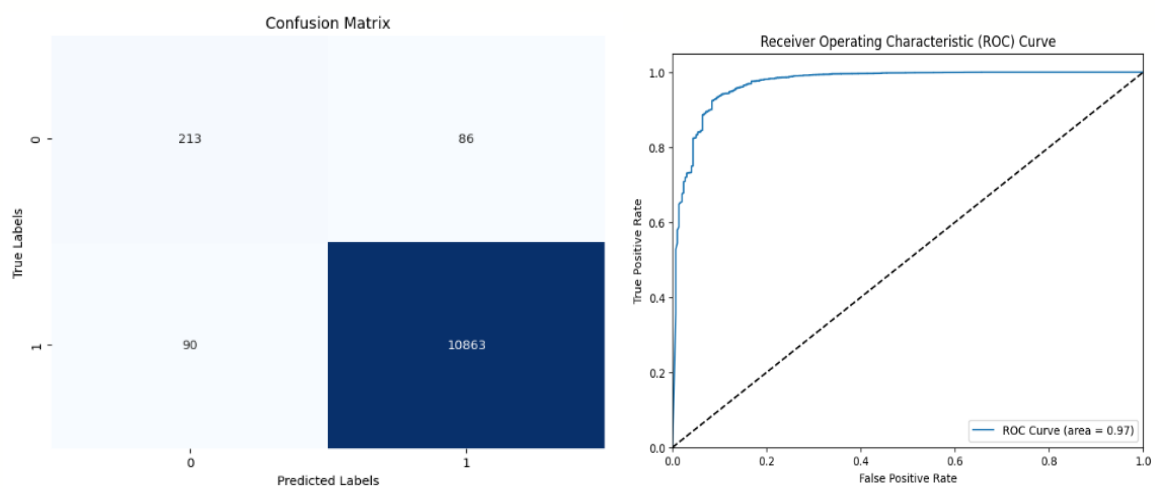


Fig 4.4: CM and the ROC-AUC Curve for Hybrid CNN + GRU.

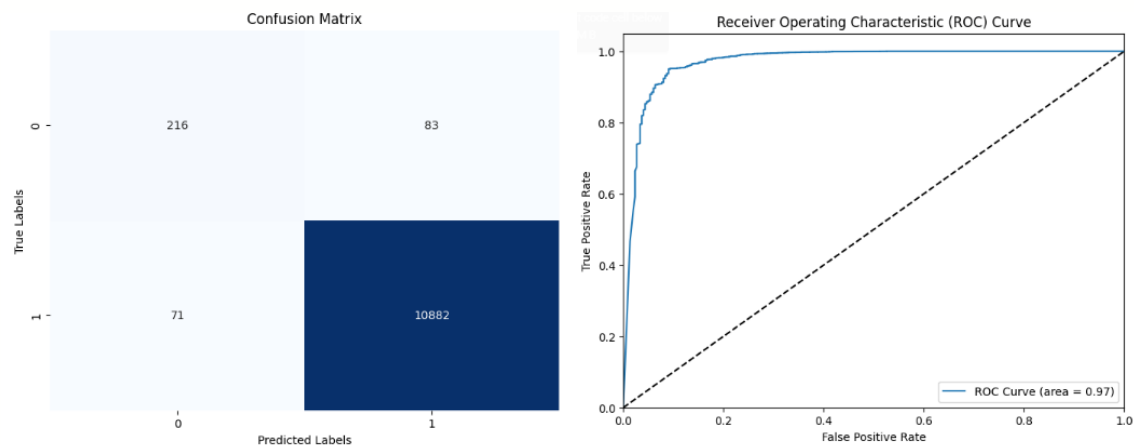


Fig 4.5: CM and the ROC-AUC Curve for CNN + LSTM + GlobalMaxPooling1D.

Explain how web-based prototypes were created, as seen in Fig. 4.7 below, by using the ML and DL passive aggressive approaches to produce the required Bangla composition of the label of the resultant data. Figures 4.7-8.9 above show the two types of bogus news that an app for a website may use to determine whether or not it is real. As is often the case, one of the more successful forms of unsupervised learning compared to diagnosis or predictions is deep learning. When compared to alternative methods, it did rather well.



Fig 4.6: Detecting bogus news in Bangla with a web prototype.



Fig 4.7: Real news detection utilizing web prototype.



Fig 4.8: Fake news detection utilizing web prototype.

4.4 Discussion

We looked at strategies for handling false information in Bengali according to this research. In an effort to close the fake news detecting gap, our work concentrated on the utilization of resource-free language, given the speed at which false information spreads. We used a variety of state-of-the-art deep learning models, including 1D CNN, LSTM, hybrid construction, and bidirectional GRU, along with a meticulously built dataset, in an attempt to improve the accuracy and reliability of automatic Bengali fake news identification. Following a comprehensive performance

evaluation, bidirectional GRU yielded remarkable outcomes concerning precision, precision, accuracy, recall, and F1 scores. The matrices of confusion and ROC-AUC curves provided further insights into the model's efficacy and selectivity, both prior to and once the data set was balanced. Apart from contributing to the expanding body of literature in the fight against misinformation, this study provides the perspectives and understanding needed to address similar problems in other languages with few resources.

Precision: Precision is the model's ability to distinguish true positives from all predicted positives. There is the following formula available:

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

Recall: The ability of the model to correctly identify each true positive amongst the anticipated positives is referred to as recall. The formula is explained as follows:

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

F1-Score: Through the combination of recall with precision, the F1 score provides a fair assessment of the model's work. To calculate it, use the following equation:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Accuracy: Accuracy is defined as the proportion of all recorded cases and correctly anticipated situations in the data set. calculated in this manner:

$$Accuracy = \frac{TP + TN}{Total\ Instances}$$

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Deep learning's ability to identify fake news in Bangla might have a big, far-reaching effect on society. The following are some instances of how society, especially profit society, could be affected by such a structure:

- **Promoting civic empowerment in Bengali:** By giving the people who speak Bangla a way to identify fake news in Bengali, the community has the ability to manage how they consume online material, which in turn promotes a feeling of technical sovereignty.
- **Scientific and technology developments:** The fundamental impetus for developing these sorts of systems is the development of deep learning, especially NLP study findings, which might result in breakthroughs across a variety of fields of study in science and technology.
- **Durability & Monetary Impacts:** Fake news may have an influence on investor confidence, the stock marketplace, and companies. Robust detection mechanisms might reduce these impacts and improve financial stability.
- **Technical Advancement:** Machine learning and the processing of natural languages have advanced with the creation of sophisticated algorithms for spotting false news in Bengali.

In conclusion, the way Bengali-speaking individuals consume, distribute, and trust information will be significantly impacted by a deep learning-based system for spotting bogus news in Bangla. In an environment that is becoming more digital and networked, preserving accurate data is essential. Because it tackles a serious issue, encourages conversation, maintains ethical journalism, and contributes to the establishment of a more reliable and educated digital environment, the Bangla fake news detection culture rooted in Dhaka has a generally beneficial impact. To guarantee that these networks are established morally, a balance must be achieved between the advantages and ethical issues, in addition to user privacy.

5.2 Impact on Environment

The creation and implementation of a system based on deep learning in Bangla for identifying bogus news may have a number of unexpected consequences for the environment. The main focus is on deception within the Bengali-speaking population, however there may be additional elements to take into account as well:

- **Resource use:** Deep learning models, particularly big ones, may need a lot of training data and might be computationally expensive. This might result in increased power use and possibly a larger carbon impact.
- **Technological Architecture:** In along with establishing and upholding the physical infrastructure required for model installation and development, it may also be essential to produce, use, and discard gadgets. This might have a bad effect on the ecosystem.
- **Situation Knowledge and Circumstances Care:** Manufacturers and organizations can reduce the possible environmental effect by taking proactive measures to mitigate it. It takes into account ecological computing techniques and the utilization of power through renewable sources.
- **Setting an assumption and collecting results:** It frequently requires fewer processing resources to draw findings from a model that has been trained than to train it. Even so, one must still take into account the resources needed for a real-time or almost real-time launch.

In general, as resolving pressing social issues is the main objective of a Bangla false news detection system, its ecological impact must be taken into account. Energy saving, software and equipment efficiency, and the technological improvement of sources of clean energy can all help to lessen the adverse environmental impacts.

5.3 Ethical Aspects

It is important to carefully consider the ethical issues surrounding the identification of false news in Bangla using DL-based techniques. Following is a list of a few of the most important ethical considerations:

Ascertain the legal and ethical accuracy of the data collected for the model's assessment and training. Be mindful of user privacy and obtain all required consents before using any data. Proceed with caution while taking action to safeguard the identity of anyone featured in news items,

especially when dealing with private or sensitive information. Recognize any possible biases in the model or data. Ensure that no individual is exposed to prejudice or unfairly granted priority due to their age, color, religion, or any other aspect of their economic status. When utilizing an online resource, users should be made aware of the presence and intended application of any technology that is embedded. Offering sign-up forms or basic alerts is one option. Make a conscious choice to maximize the transparency and comprehensibility of the framework's possibilities. Clients need to know the logic behind a certain categorization. To guarantee that the creation and application of the framework conform to ethical principles, seek counsel and supervision from ethicists, lawyers, and other pertinent stakeholders.

In summary, companies and other entities may guarantee the intelligent application of the system's Bangla fake news detector, consideration for user privacy, and enhancement of an educated and reliable news ecosystem by tackling certain ethical concerns.

5.4 Sustainability Plan

Using techniques that reduce detrimental consequences on society, the environment, and the economy is necessary to build a long-term plan for deep learning-based Bengali false news identification. The long-term planning strategy that follows is suitable for this undertaking:

- **Cost-effective Computation:** Use energy-efficient hardware and computer applications for model development and deployment. Optimize design and operations to reduce energy use.
- **Integrating energy from renewable resources with cloud computing:** Use cloud services from providers whose data centers give priority to renewable energy. Consequently, the environmental impact of developing and utilizing model software is diminished.
- **Data safety and anonymity:** Establish strong confidentiality procedures to safeguard user data. Observe privacy regulations pertaining to data and get required authorizations prior to using data.

Studying energy-saving design notions: Investigate energy-efficient design language paradigms and techniques. Seek for techniques that yield top-notch outcomes with fewer processing power consumption. The Bangla false news identification system may be developed and implemented

with a reduced ecological footprint and an effective approach to dealing with dishonesty by implementing certain sustainable approaches. This strategy is in line with the ethical and conscientious progress of AI.

CHAPTER 6

Conclusion and Future Research

6.1 Summary of the Study

We've learned a lot about this subject thanks to this inquiry. Predictors of false news are still a contentious issue. The yearly decline in patient usage of fake news detection is significantly impacted by this. Because of this, we classified the two words in the class True (1) or untrue (0) that pose a significant danger using deep learning. False news from Bangladeshi was additionally classified based on appearance using the ML model.

Our research's goal is to understand as much as we can about this subject, as we have already said. In order to do this, we combined data for four data folders—two sets of labels for genuine and fake—from online sources. Using this data, we were able to verify and hone our five techniques based on the passive aggressive algorithm's greater precision in finding the "content" of the "label" through the use of the data sets labeled as true or false. When we implement an idea for a web application, it becomes simpler for us to distinguish between fake and real news.

6.2 Conclusion

Through our research, we aimed to specifically target the spread of deceptive content in the Bangla language. Our goal is to improve fraud detection by using state-of-the-art deep learning algorithms by creating and analyzing large amounts of data. Our models' ability to identify false positives is demonstrated by the good results we observed in terms of accuracy, precision, recall, and F1 scores. However, the adventure is not over yet. We aim to add better pseudo-events to the collection of data in our next work since we recognize the importance of having equal data and because doing so will result in an improved and equitable training environment. This enhancement will improve the versatility of our models and enable practical implementations. We anticipate that as we continue to improve our methods, we will be able to identify fake news in other languages and help reduce the amount of misleading material in Bangla. We want to expand on the success of this study and provide a unique way for identifying fake news in Bangla by utilizing machine learning (ML) techniques to categorize an article's classification as True or False and developing a web-based application for the purpose of identifying false news in Bangla. People on the internet will be able to discern between real and fake news thanks to this.

6.3 Possible impacts

We believe that the outcomes of our effort differ depending on the observation location. We thus took additional effort to do our jobs accurately. We take precautions to guarantee the legitimacy of the work we do. This is why we have opted to use a digital dataset. Reducing the spread of misleading information has a big influence and might result in a public conversation that is more factual and objective. By guaranteeing consumers can rely on accurate data, helping to uncover and emphasize false news will help sustain public confidence in credible news sources. Users may assess and interact with internet material objectively by having a tool to confirm the veracity of content in Bengali. There is a significant potential effect when DL and ML identify fake news in Bangla. Not only does it address a big social issue, but it also encourages innovation, backs transparent reporting, and strives to create a more knowledgeable and reliable online community.

6.4 Implication for Further Study

It seems to indicate that there are several avenues for further research in this area, chief among them being our own. There are several suggestions from everyone for improving our work. As previously said, we have also discovered a few inaccuracies, and these faults help us to improve our research. By implementing our work and enhancing prediction results with increased opportunities to capture exceptional accuracy, we will attempt to address this issue. We intend to correct any mistakes that we find.

Further investigation might result in the creation of more precise and efficient algorithms for identifying false information in Bangla. As a result, there may be more reasons to be optimistic about the method's forecasts. We also have a few more goals in mind. There may be room for enhancement in this area of the prediction. We will include components like false news detector types in order to maximize this and help the user. We employ the ML model in conjunction with other features to construct a tool that consumers may use to predict bangla news. We think that this kind of work might improve our connections with verification groups and projects, improve the quality of the system, and support larger campaigns to combat erroneous data. For the sake of more research, we might be able to assist with real-time bogus news detection.

Reference

- [1] T. Khan, A. Michalas, and A. Akhunzada, “Fake news outbreak 2021: Can we stop the viral spread?,” *Journal of Network and Computer Applications*, vol. 190, p. 103112, Sep. 2021, doi: 10.1016/j.jnca.2021.103112.
- [2] Q. A. R. Adib, M. H. K. Mehedi, M. S. Sakib, K. K. Patwary, M. S. Hossain, and A. A. Rasel, “A Deep Hybrid Learning Approach to Detect Bangla Fake News,” *ISMSIT 2021 - 5th International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings*, pp. 442–447, 2021, doi: 10.1109/ISMSIT52890.2021.9604712.
- [3] L. Materassi, L. Solito, G. Ducci, and others, “Public sector communication facing the challenges of open government,” in *Book of Abstracts. 6th World Conference on Media and Mass Communication MEDCOM 2020+ 1*, 2021, p. 20.
- [4] M. U. Tariq, S. B. Ismail, M. Babar, and A. Ahmad, “Harnessing the power of AI: Advanced deep learning models optimization for accurate SARS-CoV-2 forecasting,” *PLoS One*, vol. 18, no. 7, p. e0287755, Jul. 2023, doi: 10.1371/JOURNAL.PONE.0287755.
- [5] M. Z. Hossain, M. A. Rahman, M. S. Islam, and S. Kar, “BanFakeNews: A Dataset for Detecting Fake News in Bangla,” *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pp. 2862–2871, Apr. 2020, Accessed: Aug. 15, 2023. [Online]. Available: <https://arxiv.org/abs/2004.08789v1>
- [6] R. I. Rasel, A. H. Zihad, N. Sultana, and M. M. Hoque, “Bangla Fake News Detection using Machine Learning, Deep Learning and Transformer Models,” *Proceedings of 2022 25th International Conference on Computer and Information Technology, ICCIT 2022*, pp. 959–964, 2022, doi: 10.1109/ICCIT57492.2022.10055592.
- [7] M. M. Hossain, Z. Awosaf, M. S. H. Prottoy, A. S. M. Alvy, and M. K. Morol, “Approaches for Improving the Performance of Fake News Detection in Bangla: Imbalance Handling and Model Stacking,” *Lecture Notes in Networks and Systems*, vol. 437, pp. 723–734, 2022, doi: 10.1007/978-981-19-2445-3_51/COVER.
- [8] M. A. Ali, M. L. Matubber, V. Sharma, and B. Balamurugan, “An Improved and Efficient Technique for detecting Bengali Fake News using Machine Learning Algorithms,” *2022 6th International Conference on Computing, Communication, Control and Automation, ICCUBE 2022*, 2022, doi: 10.1109/ICCUBE54992.2022.10011096.
- [9] F. L. Piya, R. Karim, and M. S. Arefin, “BDFN: A Bilingual Model to Detect Online Fake News Using Machine Learning Technique,” pp. 799–816, 2022, doi: 10.1007/978-981-16-5301-8_56.
- [10] S. Saha, A. Sarker, P. Chakraborty, and M. A. Yousuf, “Bengali Fake News Detection: Transfer Learning Based Technique with Masked LM Process by BERT,” *Communications in Computer and Information Science*, vol. 1670 CCIS, pp. 83–96, 2022, doi: 10.1007/978-3-031-20977-2_7/COVER.
- [11] M. Iftikhar and A. Ali, “Fake News Detection using Machine Learning,” *3rd IEEE International Conference on Artificial Intelligence, ICAI 2023*, pp. 103–108, 2023, doi: 10.1109/ICAI58407.2023.10136676.
- [12] N. L. Siva Rama Krishna and M. Adimoolam, “Fake News Detection System using Logistic Regression and Compare Textual Property with Support Vector Machine Algorithm,” *International Conference on Sustainable Computing and*

Data Communication Systems, ICSCDS 2022 - Proceedings, pp. 48–53, 2022, doi: 10.1109/ICSCDS53736.2022.9760768.

- [13] T. Al Mahmud, S. Sultana, and A. Mondal, “A New Technique to Classification of Bengali News Grounded on ML and DL Models Digital Evaluation of Broad Question Answer Script View project Blockchain View project A New Technique to Classification of Bengali News Grounded on ML and DL Models,” *Article in International Journal of Computer Applications*, vol. 185, no. 18, pp. 975–8887, 2023, doi: 10.5120/ijca2023922897.
- [14] E. Hossain, Md. Nadim Kaysar, A. Z. Md. Jalal Uddin Joy, Md. Mizanur Rahman, and W. Rahman, “A Study Towards Bangla Fake News Detection Using Machine Learning and Deep Learning,” pp. 79–95, 2022, doi: 10.1007/978-981-16-5157-1_7.
- [15] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, “Fake News Detection Using Machine Learning Ensemble Methods,” *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8885861.
- [16] N. Ahuja and S. Kumar, “S-HAN: Hierarchical Attention Networks with Stacked Gated Recurrent Unit for Fake News Detection,” *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, pp. 873–877, Jun. 2020, doi: 10.1109/ICRITO48877.2020.9197914.
- [17] R. K. Kaliyar, A. Goswami, and P. Narang, “FakeBERT: Fake news detection in social media with a BERT-based deep learning approach,” *Multimed Tools Appl*, vol. 80, no. 8, pp. 11765–11788, Mar. 2021, doi: 10.1007/S11042-020-10183-2/TABLES/22.
- [18] A. Yasin, M. J. Awan, H. Muhammad, and F. Shehzad, “Fake News Classification Bimodal using Convolutional Neural Network and Long Short-Term Memory,” *Article in International Journal of Emerging Technologies in Learning (iJET)*, 2020, Accessed: Aug. 15, 2023. [Online]. Available: <https://www.researchgate.net/publication/347886404>
- [19] “BanFakeNews | Kaggle.” <https://www.kaggle.com/datasets/cryptexcode/banfakenews> (accessed Aug. 15, 2023).
- [20] M. K. Dahouda and I. Joe, “A Deep-Learned Embedding Technique for Categorical Features Encoding,” *IEEE Access*, vol. 9, pp. 114381–114391, 2021, doi: 10.1109/ACCESS.2021.3104357.
- [21] B. Zhong, X. Pan, P. E. D. Love, J. Sun, and C. Tao, “Hazard analysis: A deep learning and text mining framework for accident prevention,” *Advanced Engineering Informatics*, vol. 46, p. 101152, Oct. 2020, doi: 10.1016/J.AEI.2020.101152.
- [22] “stopwords-iso/stopwords-bn: Bengali stopwords collection.” <https://github.com/stopwords-iso/stopwords-bn> (accessed Aug. 15, 2023).
- [23] T. Chauhan and H. Palivela, “Optimization and improvement of fake news detection using deep learning approaches for societal benefit,” *International Journal of Information Management Data Insights*, vol. 1, no. 2, p. 100051, Nov. 2021, doi: 10.1016/J.JJIMEI.2021.100051.
- [24] W. C. Lin and C. Busso, “Chunk-Level Speech Emotion Recognition: A General Framework of Sequence-to-One Dynamic Temporal Modeling,” *IEEE Trans Affect Comput*, vol. 14, no. 2, pp. 1215–1227, Apr. 2023, doi: 10.1109/TAFFC.2021.3083821.

- [25] T. Islam, M. A. Hosen, A. Mony, M. T. Hasan, I. Jahan, and A. Kundu, “A Proposed Bi-LSTM Method to Fake News Detection,” *2022 International Conference for Advancement in Technology, ICONAT 2022*, 2022, doi: 10.1109/ICONAT53423.2022.9725937.
- [26] Md. A. Sheakh, Mst. Sazia Tahosin, M. M. Hasan, T. Islam, O. Islam, and M. M. Rana, “Child and Maternal Mortality Risk Factor Analysis Using Machine Learning Approaches,” in *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*, IEEE, May 2023, pp. 1–6. doi: 10.1109/ISDFS58141.2023.10131826.

USING_DEEP_LEARNING_TECHNIQUES_TO_IDENTIFY_FALSE_...

ORIGINALITY REPORT

18%	16%	9%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to Jacksonville University Student Paper	1%
3	Submitted to BRAC University Student Paper	1%
4	arxiv.org Internet Source	1%
5	dokumen.pub Internet Source	1%
6	Submitted to George Bush High School Student Paper	1%
7	thesai.org Internet Source	1%
8	Arwa A. Al Shamsi, Sherief Abdallah. "Ensemble Stacking Model for Sentiment Analysis of Emirati and Arabic Dialects", Journal of King Saud University - Computer and Information Sciences, 2023 Publication	<1%