

**FAKE NEWS DETECTION ON BENGALI LINGUISTIC USING DEEP
LEARNING APPROACH**

BY

Shihab Shahariar

ID: 201-15-14105

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

Supervised By

Ms. Shayla Sharmin

Senior Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

Mr. Tanvirul Islam

Lecturer

Department of CSE

Daffodil International University

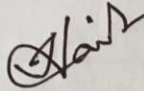


**DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2024**

APPROVAL

This Project titled “**FAKE NEWS DETECTION ON BENGALI LINGUISTIC USING DEEP LEARNING APPROACH**”, submitted by Shihab Shahariar, ID: 201-15-14105 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 25 January, 2024.

BOARD OF EXAMINERS

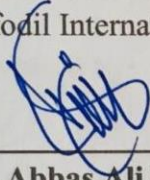


Chairman

Dr. Sheak Rashed Haider Noori

Professor & Head

Department of Computer Science and Engineering
Daffodil International University

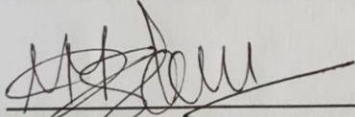


Internal Examiner-1

Md. Abbas Ali Khan

Assistant Professor

Department of Computer Science and Engineering
Daffodil International University



Internal Examiner-2

Mohammad Monirul Islam

Assistant Professor

Department of Computer Science and Engineering
Daffodil International University



External Examiner-1

Dr. Md. Arshad Ali

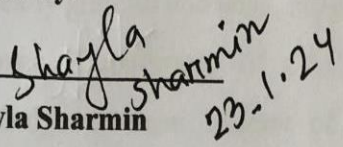
Professor

Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology
University

DECLARATION

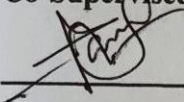
I hereby declare that this project has been done by us under the supervision of **Ms. Shayla Sharmin, Senior Lecturer, Department of CSE, Daffodil International University**. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree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

Supervised by:



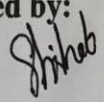
Ms. Shayla Sharmin
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Tanvirul Islam
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Shihab Shahariar
ID: 201-15-14105
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

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

I am really grateful and wish, my profound indebtedness to **Ms. Shayla Sharmin, Senior Lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Natural language processing (NLP)*” to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Sheak Rashed Haider Noori, Professor & Head**, Department of CSE and **Ms. Shayla Sharmin, Senior Lecturer**, Department of CSE, for their kind help to finish my project and also to other faculty members and the staff of CSE department of Daffodil International University.

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

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

ABSTRACT

A vast quantity of data and information are available on the internet. Because the internet is so widely available and has resulted in a tremendous growth in the number of online news, people are interested in reading news from online news portals. Online news portals include things like Facebook, Twitter, WhatsApp, Telegram, Instagram, blogs, and more. Both the quantity of news-on-news websites and the number of readers is increasing. But how real is online news today is a matter of thought. A huge amount of fake news is being spread in newspapers and online due to various yellow journalists. Which is having an adverse effect on society. As a result, there are many kinds of instability, bad politics, etc. problems are being created in the country. If this situation continues, our country and society will go to hell. The only solution is to ensure that yellow journalists do not spread fake news. But despite all the vigilance, fake news will spread. We can solve this by using artificial intelligence, for example, by employing various machine learning and deep learning algorithms, we can identify bogus news and take precautions against it. In this paper, fake news is detected using 4 deep learning algorithms like RNN, LSTM, BiLSTM, GRU model and 1 machine learning algorithm BERT model. RNN has an accuracy of 94.58%, LSTM has an accuracy of 92.84%, BiLSTM has an accuracy of 94.29%, GRU has an accuracy of 93.22% and BERT has an accuracy of 95%. The BERT model has the highest accuracy of 95% among all models.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
List of figures	ix
List of tables	xi
CHAPTER	
CHAPTER 1: INTRODUCTION	1-5
1.1. Introduction	1
1.2. Motivation	2
1.3. Relational of the Study	3
1.4. Research Questions	3
1.5. Expected Outcome	4
1.6. Report Layout	4
CHAPTER 2: BACKGROUND	5-11
2.1. Terminology	5
2.2. Related work	5
2.3. Comparative Analysis and Summary	8
2.4. Scope of the Problem	10
2.5. Challenges	11

CHAPTER 3: RESEARCH METHODOLOGY	12-40
3.1. Data Pre-Processing	12
3.2. Dataset cleaning	12
3.3. Statical Analysis	13
3.4. Design Approach	13
3.5. Dataset Description	14
3.6. Label Encoding	17
3.7. Tokenizer	17
3.8. Recurrent Neural Network (RNN)	18
3.9. Long Short-Term Memory (LSTM)	21
3.10. Bidirectional Long Short-Term Memory (Bi-LSTM)	24
3.11. Gated Recurrent Units (GRU)	28
3.12. Bidirectional Encoder Representations from Transforms (BERT)	31
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	40-41
4.1. Discussion	40
4.2. Experimental Results and Analysis	41
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	42-44
5.1. Impact on Society	42
5.2. Impact on Environment	42
5.3. Ethical Aspects	43
5.4. Sustainability	43

CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLEMENTATION FOR FUTURE RESEARCH	44-46
6.1. Summary of the Study	44
6.2. Conclusions	44
6.3. Implication for Further Study	45
REFERENCE	46

LIST OF FIGURES

FIGURES	PAGE NO
Figure 1: Architecture of Working Process	14
Figure 2: Dataset Distribution	15
Figure 3: Dataset Statistics	16
Figure 4: Length-frequency distribution	17
Figure 5: Epochs vs training and validation accuracy plot for RNN	19
Figure 6: Epochs vs training and validation loss plot for RNN	20
Figure 7: Confusion Matrix of RNN	21
Figure 8: Epochs vs training and validation accuracy plot for LSTM	22
Figure 9: Epochs vs training and validation loss plot for LSTM	23
Figure 10: Confusion matrix for LSTM	24
Figure 11: Epochs vs training and validation accuracy plot for Bi-LSTM	25
Figure 12: Epochs vs training and validation loss plot for Bi-LSTM	26
Figure 13: Confusion matrix for Bi-LSTM	27
Figure 14: Epochs vs training and validation accuracy plot for GRU	29
Figure 15: Epochs vs training and validation loss plot for GRU	30
Figure 16: Confusion matrix for GRU	31
Figure 17: Authentic Headlines Length TOP 100	32
Figure 18: Authentic Content Length TOP 100	32
Figure 19: Authentic Headline Length TOP 100	33
Figure 20: Authentic Content Headline Length TOP 100	33
Figure 21: Fake Headlines Length TOP 100	34
Figure 22: Fake Content Headline Length TOP 100	34
Figure 23: Fake Headline Length TOP 100	35
Figure 24: Fake Content Headline Length TOP 100	35
Figure 25: Word Cloud Image for Authentic Headlines	36
Figure 26: Word Cloud Image for Fake Headlines	37
Figure 27: Stop Words for Authentic Headlines	38
Figure 28: Stop Words for Fake Headlines	38
Figure 29: Epochs vs Training & Validation Loss Plot for BERT	40

LIST OF TABLES

TABLES	PAGE NO
Table 1: Summary of Related Works	8
Table 2: Cleaned data from dataset	12
Table 3: Data Statistics	15
Table 4: Validation and Training accuracy table for RNN	18
Table 5: Validation and Training loss for RNN	19
Table 6: Classification report of RNN	20
Table 7: Test RNN algorithm using dummy news	21
Table 8: Validation and Training accuracy table for LSTM	22
Table 9: Validation and Training loss table for LSTM	23
Table 10: Classification report for LSTM	23
Table 11: Test LSTM algorithm using dummy news	24
Table 12: Validation and Training accuracy table for Bi-LSTM	25
Table 13: Validation and Training loss table for Bi-LSTM	26
Table 14: Classification report for Bi-LSTM	27
Table 15: Test Bi-LSTM algorithm using dummy news	28
Table 16: Validation and Training accuracy table for GRU	28
Table 17: Validation and Training loss table for GRU	29
Table 18: Classification report for GRU	30
Table 19: Test GRU algorithm using dummy news	31
Table 20: Validation loss, Training loss and Accuracy for BERT	39
Table 21: Classification report of BERT algorithm	40
Table 22: Classifiers Description	41
Table 23: Classifiers accuracy, recall and precision	42

CHAPTER 1

INTRODUCTION

1.1. Introduction

The primary information source and essential component of peoples' life is the Internet. Globally, the number of online news sources is rapidly expanding, and individuals are increasingly interested in reading daily news portals as a result of internet accessibility. Bengali headlines and hourly news updates are available on thousands of portals. Nowadays, in the era of digitalization, the majority of people consume news online rather than in newspapers. The use of online portals, Twitter, Facebook, blogs, and other apps is increasing these days, which is why there is a lot of information available on websites and a growing quantity of usage of the internet. In other words, a large portion of the global population now exclusively relies on Internet news. Along with producing and accessing a lot of information every day, people are also using handheld multimedia devices and high-speed Internet. As of the end of January 2018, there were around 80.83 million internet users in Bangladesh, of which approximately 30 million used social media. Since the paper boom has diminished, there has been a daily growth in the creation and consumption of online news, with an emphasis on the Internet. Rather than disseminating their news, many news organizations seem to be producing and uploading it to the internet. E-news is the term used to describe news that is published and made available via the Internet. The viewership of this news is growing daily because to internet user scholarship. As a result, a wide variety of news are being entered into the website's database. In emerging nations like India, Bangladesh, and Pakistan, news plays a critical role in disseminating knowledge and raising public awareness of events in neighboring countries.

Bengali is our mother tongue as citizens of Bangladesh, and many martyrs gave their lives in 1952 to protect it. In Bangladesh and some parts of India, Bengali is a commonly spoken language. Currently, 228 million people speak Bengali on a daily basis, of which 37 million are non-native speakers. Based on a global census of speakers, Bengali is the eighth most spoken language in the world. Recent decades have demonstrated that news has a rapid impact on the public, both positively and negatively. Numerous studies demonstrate that readers are negatively impacted by the bad news. It appears that more

negative news than positive news is being released in the majority of the videos. People are affected by these occurrences or impacts in both a bodily and mental way. These psychological impacts include anxiety or pessimism, despair, lack of focus, tension, initiative, fear, and so forth.

One of the foundational ideas of information technology is effective information retrieval. News headlines are a more generalized kind of text content. Online sources of news include those regarding computers, social sciences, music, politics, Hollywood, Bollywood, sports, and entertainment. On the internet, users may find and see any kind of news. Users may quickly search for and see news based on their needs by using news headlines. And in addition to watching, the news can be understood whether it is real or fake thanks to the current artificial intelligence.

1.2. Motivation

The field of automatic text categorization is still developing. Text categorization is becoming more and more necessary to handle the world's expanding digital data, which is developing at an incredible rate. By categorizing data and applying various data mining algorithms for text classification, false news detection, etc., fake news is detected using machine learning algorithms. Text classification is used in applications including corporate analytics, spam screening, and content labelling. Websites in the Bengali language contain vast amounts of material, making it challenging to locate reliable information. However, classifying readers and hash-tagging keywords become crucial tasks if you wish to post on a forum. Because of the use of hash tagging, it has become an important task to find out whether the news goal of the hash tagging keyword is authentic or not. If not, people will use hash tagging but will fail to hash tag fake news or hash tag authentic news. Because there has never been any work done on any application platform to detect fake news using text categorization—which is a major issue—fake news detection using status categorization has grown in importance for the Bengali language. Once more, because text mining is becoming more and more necessary, academics from Bangladesh and India are concentrating on creating text mining applications. There are a lot of Bengali-language text mining and sentiment analysis studies that demonstrate effective removal. Before most people read anything in the news or online, Regarding the news, people create a judgement on how legitimate, logical, or fraudulent the claims are, as well as whether the news is true or fake. NLP can

address any kind of linguistic issue in order to identify Bengali false news. The best algorithms for understanding natural language processing (NLP) problems are machine learning ones since they try to convey ideas from problems using human language. Three types of machine learning exist: supervised, unsupervised, and semi-supervised. Semi-supervised, unsupervised, and supervised learning. Input and output are necessary for supervised learning, in addition to flat data. Input and output are necessary for unsupervised learning, in addition to layer-free data. Semi-supervised learning is simply a combination of supervised and unsupervised learning using both labelled and unlabeled data. But sorting through the deluge of information and news to decide what is true and what is fraudulent takes time, effort, and skill. These time-consuming, tough, and complex algorithms are defeated by machine learning algorithms. Fake news identification has become more popular in recent years due to advancements in machine learning. The preference of users for websites that consistently offer breaking news and updated news is another feature of e-news in Bangladesh. You must thus start by visiting the well-known and frequented websites. Prothom Alo, Bangladesh Protidin, Nayadiganta, Jugantor, Samakal, and other sites are in the top five. It is evident from examining Google Trends data that fewer people are reading "Daily News" online on a daily basis.

1.3. Relational of the Study

To determine whether a news article belongs in which category, our study paper is split into two sections. According to our study, news may be classified as either authentic or phony. We discovered other such efforts; they classified the news into similar categories and used various algorithms, such as deep learning and machine learning, to determine the news' safety.

1.4. Research Questions

This research may include a wide range of question kinds. As an illustration,

- What is the aim of this study?
- Why is this study being conducted?
- Can this research help us in any way?
- Which path can this research help us with?
- What are this study's findings?

- What is the study's conclusion?

1.5. Expected Outcome

News is a kind of information that may be used for future research projects and to have a true understanding of any kind of activity. Since the goal of our research is to identify false news in various news portals, after it is finished, I will be able to determine the news's legitimacy using any type of data. Additionally, a great deal of genuine data may be prevented from being lost since a great deal of unclassified data that is useless or that hasn't been verified as real or false can be verified as real or fake and turned into useful data.

1.6. Report Layout

But since the news portals are not effectively classifying, organizing, and evaluating the information they gather from social media and online sources, they are throwing away a lot of potentially useful material. If the data can be automatically classified using machine learning (ML), deep learning (DL), and natural language processing (NLP) in a way that is quicker, more flexible, more reliable, and more economical, then this classification might be a significant potential solution. The essay's remaining section is organized as follows: Part 2: Context; Part 3: Approach; Part 4: Findings and discussion from the experiment; and Part 5: Consequences for the environment, society, and sustainability. Section 6 will serve as the paper's final conclusion.

CHAPTER 2

BACKGROUND

2.1. Terminology

There are happenings everywhere as the globe keeps growing. Due to the internet, the aforementioned occurrences are also becoming more widespread on social media. In this sense, the volume of news keeps growing over time. To what extent, however, phony or authentic news is undetected? As a result, a significant amount of internet data is lost and is never recovered. As a result, the data is categorized using various machine learning algorithms to determine whether or not all of the documents or data are pertinent or valuable. in order for the facts of today to be useful for tomorrow.

2.2. Related Works

Kingaonkar et al. [1] have proposed a study article in which various forms of authentic false news are gathered from social media platforms or online news sites and categorized using various machine learning algorithms. The dataset comprises columns 5–10 and 2000 news items. The support vector machine (SVM) algorithm is one of the machine learning algorithms. 99.90% of the time, the Support Vector Machine (SVM) algorithm is accurate. In their research work, Gaikwad et al. [2] provide a voting classifier set and feature extraction and selection method for the detection of false news. Following the dataset's preprocessing, two feature extraction methods—word frequency-inverse document frequency approach and bag of words—were used. Additionally, a few machine learning algorithms—like SVM, LR, and NB—are used. Adedoyin et al.'s study article [3] aims to identify false news and authentic news by gathering information from social media and online news sites. Finding the optimal performance model through the use of machine learning algorithms to identify false news is the primary task at hand. Deep learning rnn models, Multinomial NB, Random Forest, SVM, and Logistic Regression are utilized in machine learning. In 99% of machine learning cases, it has been observed that the SVM method yields the highest accuracy. A study by Choudhary et al. [4] examines how to identify false news by gathering both actual and fraudulent news from a variety of online news portals and social media platforms, including Facebook, LinkedIn, Twitter, and WhatsApp. G. Senthilkumar et al. [5] propose a

research paper mainly deals with identifying fake news through natural language processing of machine learning. Fake news detection is done using Naïve Bayes, Convolutional Neural Network, LSTM, Neural Network, SVM, and N-Gram Analysis as a machine learning algorithm. Three machine learning algorithms, namely Passive Aggressive Classifier, Naive Bayes, and Random Forest, are employed in this research study. As can be seen from the three accuracies, the passive aggressive classifier approach has the highest accuracy (87%). The study article proposed by Rahman et al. [6] focuses on the identification and dissemination of fake news on online portals, social media, and microblogging networks. Four machine learning algorithms—logistic regression (LR), decision tree (DT), k-nearest neighbors (KNN) and naive bayes (NB)—and two deep learning algorithms—long short-term memory (LSTM) and bidirectional long-short-term memory (Bi-LSTM) method—are covered. The machine learning algorithms that produced the greatest accuracy results were Logistic Regression (96%), followed by Bi-LSTM (99%) in deep learning. A research study is proposed by Kulkarni et al. [7] to investigate if a machine learning model for detecting false news has been constructed using various techniques. Typically, news from different news sources, news portals, social media, etc. is utilized to populate the model. The machine learning algorithms employed in this context include the Gradient Booster Algorithm, Random Forest, Decision Tree, Logistic Regression, and KNN. With an accuracy of 85%, the logistic regression method is the most accurate machine learning technique. In their study work, Choudhury et al. [8] analyze how false news identification has been accomplished primarily by data collection from several sources, such as social media and online news portals. Here, a variety of machine learning methods are applied to identify false news. SVM, Random Forest, Naïve Bayes, and Logistic Regression are a few of the machine learning algorithms that are employed. Three areas comprise SVM: false information, phony job listings, and phony news. SVM and RF both get the best accuracy of 61% among the four algorithms on the LIAR dataset; on the false job posting dataset, both methods provide the highest accuracy of 97%. In essence, the study work by GOMATHY et al. [9] uses natural language processing (NLP) to distinguish between authentic and fraudulent news. After running the gathered news via a machine learning algorithm, the frequency of the news was determined. News from many internet news sites, including Facebook and Twitter, has been gathered and examined. A study work that focuses on identifying and detecting false news linked to medicine is proposed by Murugesan et al. [10]. Here, five machine learning algorithms—KNN, Naive Bayes, Support Vector

Machine, BERT, and Decision Tree—are employed. With the Adaboost & Decision Tree method, the greatest accuracy of 98.5% was discovered out of all the algorithms. In a study article by Khanam et al. [11], they propose using several libraries such as Python Skit-Learn, NLP, etc. to identify and analyze fake news, classifying it as real or false using machine learning. Machine learning methods including XGboost, Random Forests, Naive Bayes, K-Nearest Neighbors (KNN), Decision Trees, and Support Vector Machines (SVMs) are utilized. At 75%, XGboost offered the best accuracy. In essence, a research study by Kushwaha et al. [12] uses machine learning algorithms to identify bogus news. In this case, three machine learning techniques are used for the analysis: Random Forest Classification, Naïve Bayes, and Logistic Regression. With an accuracy of 65%, the logistic regression method yielded the best results of all the techniques. In their proposed research study, Ngada et al. [13] primarily investigate machine learning methods for the purpose of detecting bogus news. Six machine learning algorithms—AdaBoost, Decision Tree, K-Nearest Neighbors, Random Forest, Support Vector Machine, and XGBoost—are used to analyze bogus news identification in this case. With an accuracy of 99.4%, the support vector machine classifier yielded the highest accuracy of all the algorithms. Lakshmanarao et al. [14] offer an analysis of a research paper that is helpful for investigating methods of detecting fake news using machine learning algorithms, particularly those related to Natural Language Processing (NLP). Four machine learning algorithms—SVM, K-Nearest Neighbors, Decision Tree, and Random Forest—are used to analyze the identification of bogus news in this case. With an accuracy of 90.7%, the Random Forest classifier provided the best results of all the methods. A study article is proposed by Lasotte et al. [15] to identify machine learning algorithms that can more accurately anticipate the identification of false news. Six machine learning techniques are utilized in this case for prediction analysis in the detection of false news, including Naïve Bayes, SVM, Random Forest, Logistic Regression, hard voting, and soft voting. With an accuracy of 93%, the soft voting algorithm yielded the best results of all the methods. Sultana et al.'s research work [16] analyses a model for artificial intelligence, or machine learning, -based false news identification. This article describes the use of eight machine learning algorithms, such as Multinomial NB (MNB), Strategic Relapse (LR), Choice Tree Arrangement (DTC), Inclination Supporting Classifier (GBC), Arbitrary Backwoods Classifier (RFC), Direct SVC (SVC), Inactive Forceful Classifier (Dad), and K Neighbors Classifier (KNC), in the detection of fake news. With an accuracy of 96%, the Support Vector Classifier

(Straight SVC) method has the best accuracy out of all of the algorithms. In a research publication, K. Nath et al. [17] examine the effectiveness of machine learning and deep learning models for identifying false news. In this work, four data sets total have been used to apply various models employing artificial intelligence. To identify false news, three classification algorithms—TF-IDF, BOW, and Word2Vec—are employed in this instance. The algorithms that produced the best accuracy across all categories were Random Forest (RF), Logistic Regression (LR), Support Vector Classifier (SVC), and TF-IDF. In Bag-of-Words (BOW), Random Forest (RF) yielded the highest accuracy. offered, with Word2Vec's Multilayer Perceptron (MP) method yielding the best accuracy. In a research study, Md. Y. Tohabar et al. [18] essentially provide a machine learning-based artificial intelligence (AI) model for detecting bogus news. These machine learning techniques may be used to identify false news. With an accuracy of 73.20%, the support vector machine (SVM) method has the best accuracy of all the algorithms. In a study article that they suggest, S. Pandey et al. [19] analyze a model for detecting false news by employing artificial intelligence, or machine learning techniques. These machine learning methods (89.98% for KNN, 90.46% for Logistic Regression, 86.89% for Naïve Bayes, 73.33% for Decision Tree, and 89.33% for SVM) can identify bogus news. With an accuracy of 90.46%, the logistic regression method offers the best accuracy of all the algorithms. In a research study proposed by A. Santhosh Kumar et al. [20], a false news detection algorithm utilizing artificial intelligence, or machine learning techniques, is employed. Fake news may be identified using seven machine learning algorithms, including CNN, LSTM-ultimate, LSTM-Avg, BERT, Naïve Bayes, and SVM. These algorithms have been used to identify and analyze fake news.

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	Shagun Kingaonkar, Ajinkya Bawane, Ruchi Rana, Janhavi Thool, Nikita Kale (2023)	Support Vector Machine	SVM	99.90%
2	Tejaswi Gaikwad, Bhaskar Rajale, Prasad Bhosale,	SVM, LR, NB	-	-

	Swapnil Vedpathak, Mrs. S.S. Adagale (2022)			
3	Brindha Mariyappan, Festus Fatai Adedoyin (2022)	ML: KNN, Multinomial NB, Logistic Regression, Random Forest, SVM DL: RNN	SVM	99%
4	Murari Choudhary, Shashank Jha, Prashant, Deepika Saxena, Ashutosh Kumar Singh (2021)	Neural Network, Naïve Bayes, SVM, N-Gram Analysis, CNN, LSTM	-	-
5	G. Senthilkumar, D. Ashok Kumar (2023)	Random Forest, Naive Bayes and Passive Aggressive Classifier	Passive Aggressive Classifier	87%
6	Mahfujur Rahman, Mehedi Hasan, Md Masum Billah, and Rukaiya Jahan Sajuti (2022)	ML: NB, KNN, DT, and LR. DL: BiLSTM, or LSTM.	ML: Logistic Regression DL: BiLSTM	ML: 96% DL: 99%
7	Prasad Kulkarni, Suyash Karwande, Rhucha Keskar, PrashantKale and Sumitra Iyer (2021)	GB Algorithm, RF, DT, LR, and KNN	Logistic Regression	85%
8	Deepjyoti Choudhury, Tapodhir Acharjee (2022)	SVM, NB, RF and LR.	SVM & RF	97%
9	Dr. C K GOMATHY, Ms. C.V.S.VASAVI, Mr. D.Y.V.RAJESH, Ms. A.SRIJA (2021)	Natural Language Processing (NLP)		
10	Sudhakar Murugesan, Kaliyamurthi Pachamuthu (2022)	DT, BERT, NB, SVM, and KNN	Decision Tree	98.5%
11	Z Khanam, B N Alwasel, H Sirafi and M Rashid (2021)	DT, SVMs, XGboost, RF, NB, and KNN	XGboost	75%
12	Nidhi Singh Kushwaha, Pawan Singh (2022)	RF, NB, and LR	Logistic Regression	65%
13	Okuhle Ngada, Bertram Haskins (2020)	AdaBoost, XGBoost, RF, KNN, DT, SVM	Support Vector Machine	99.4%
14	A.Lakshmanarao, Y.Swathi, T. Srinivasa Ravi Kiran (2019)	RF, DT, KNN and SVM	Random Forest	90.7%

15	Y. B. Lasotte, E. J. Garba, Y. M. Malgwi, and M. A. Buhari (2022)	Hard and soft voting, RF, SVM, LR, NB, RF	Soft Voting Algorithm	93%
16	Ragia Sultana, Md. Khaled Hassan, Md. Rakibul Hassan, Saifur Rahaman Sourav, Md Abu Huraira and Shamim Ahmed (2022)	Direct SVC, Inactive Forceful Classifier, Arbitrary Backwoods Classifier, Inclination Supporting Classifier, KNN, Strategic Relapse, and Multinomial NB	Support Vector Classifier (Straight SVC)	96%
17	Keshav Nath, Priyansh Soni, Anjum, Aman Ahuja, Rahul Katarya (2021)	RF, LR, SVC, MP	TF-IDF: RF, LR, SVC BOW: RF Word2Vec: Multilayer Perceptron (MP)	-
18	Md. Yasmi Tohabar, Nahiyah Nasrah, Asif Mohammed Samir (2021)	SVM	Support vector machine (SVM)	73.20%
19	Shalini Pandey, Sankeerthi Prabhakaran, N V Subba Reddy and Dinesh Acharya (2022)	KNN, LoLR, NB, DT and SVM	Logistic Regression	90.46%
20	A Santhosh Kumar, P Kalpana, K Sharan Athithya, V Sri Ajay Sundar (2021)	BERT, LSTM-ultimate, LSTM-Avg, NB, SVM and CNN	-	-

2.4. Scope of the Problem

The world is changing over time, with various things happening at different times. Once more, social media sites like Facebook, Twitter, WhatsApp, Viber, Pinterest, Telegram, Messenger, Google, YouTube, and so forth have made them available online. The quantity of news is growing daily as a result. News is increasing in proportion to the amount of information or documents available online. But failing to identify which documents or data belong in which category (genuine or phony), that is, failing to determine whether the information is true or fraudulent. This results in the loss of a lot of internet data that may be useful later. To make all these data or documents practical,

machine learning and deep learning employ a range of algorithms, such as logistic regression, linear regression, random forest classifier, decision tree, naïve bias, SVM, CNN, RNN, ANN, LSTM, and others. Documents that are deemed pertinent or helpful are noted. Consequently, there are several uses for papers or data that may be found online.

2.5. Challenges

The amount of news that is available online is growing every day in the modern world; within a few days, there will be more useless info than useful data. The inability to determine whether the data or documents are authentic or fraudulent—that is, whether data belongs in which category—is the cause of their unitability. As a result, a significant amount of internet data is lost and is never recovered. We will have a lot of issues later on if we don't make all of this data more usable. In the future, utilizing data to make changes will take a lot of time and professional labor. It will be a waste of money and effort. Therefore, the data will be beneficial and free from issues in the future if it is determined that the data that is now being received—that is, the data that is accessible on social media or other platforms—is true or phony.

CHAPTER 3

RESEARCH METHODOLOGY

3.1. Data Pre-Processing

The information gathered via the internet is primarily unstructured. Therefore, processing the data is required once it has been collected. Due to the fact that data gathered from websites, blogs, and social media platforms sometimes contains null values, redundant data, and other errors. The dataset is cleaned, low length data is eliminated, unique data is eliminated, duplicate data is eliminated, and the dataset is converted into intelligible data. In addition, several dataset features have been extracted, such as word lists, sorted word lists, documents per class, total words per class, unique words per class, dataset splitting, etc. As a result, we must preprocess our dataset using data techniques. To use the pre-processing approach, take the actions listed below:

- Take out hashtags, screen names, and URLs.
- Eliminate all zero values, punctuation, symbols, emojis, and integers.
- Eliminate any superfluous symbols and retweets.
- Eliminate all short data.

3.2. Dataset Cleaning

There are two categories in the dataset: authentic and false. The maximum number of articles is in the real class. All emojis, symbols, and punctuation from the two data set categories, such as ((), \}, [], /, " ", " :,!,?) Cleaning was accomplished by taking out, etc. Additionally, the data collection was cleaned by deleting any short-lived documents that were found. There were 10321 papers left after 186 little ones were eliminated.

Table-2: Cleaned data from dataset

No	Original Text	Label	Cleaned Text
1	নাইক্ষ্যংছড়িতে 'বন্দুকযুতে' ডাকাে ডনহে।	Real	নাইক্ষ্যংছ ডি তে বন্দুকযু তে ডাকাে ডনহে
2	চাড়বর ভড়েি পরীক্ষায় জাড়িয়াড়ে: আটক ২।	Real	চাড়বর ভড়েি পরীক্ষায় জাড়িয়াড়ে আটক ২
3	ওস্তাতের ডনতেিতে আইএসতর গাড়ি তেই: তজাকাই িামা।	Fake	ওস্তাতের ডনতেিতে আইএসতর গাড়ি তেই তজাকাই িামা
4	তেম করতি বািতব ওজন!	Fake	তেম করতি বািতব ওজন

3.3. Statistical Analysis

- There are 10507 total data points in the dataset.
- Dataset retains six columns.
- There are 10321 original data in the collection, of which 8043 are genuine and 2278 are fraudulent.
- There are 7430 training data points in all in the dataset.
- There are 1033 testing data points in the dataset.
- There are 1858 validation data in all in the dataset.
- Real and fake labels are divided into two groups.

3.4. Design Approach

We employed both supervised and unsupervised learning techniques since the data presents a multivariate classification challenge. This study employs deep learning techniques such as RNN, LSTM, BiLSTM, GRU, and BERT algorithms. For every model, a confusion matrix, performance and accuracy predictions, and an analysis of the outcomes were made. Figure displays the overall system architecture diagram:

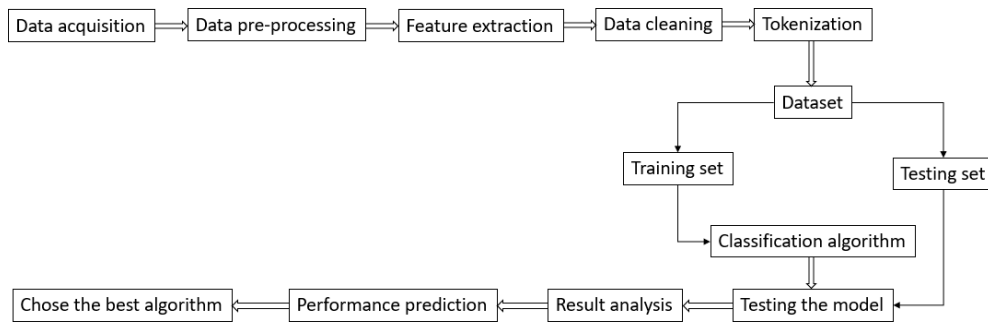


Figure-1: Architecture of Working Process

3.5. Dataset Description

Six columns make up the primary dataset utilized in this work: domain, date, category, title, content, and label. Two categories comprise the label, namely- original and fake. Besides, there are 10507 rows of data in the dataset. By preprocessing or cleaning the dataset, we used a fundamental partitioning of the dataset consisting of 10321 original documents and removed 186 documents, 7430 samples for training, 1033 samples for testing and 1858 samples for validation. After data preprocessing, out of total 10321 data, 8043 real data and 2278 fake data were found.

Distribute the dataset when it has been prepared. The collection has 10321 documents in total. The 10321 papers are categorized as true or phony, with the real documents having a higher count than the bogus documents. A graph of the data distribution is shown below:

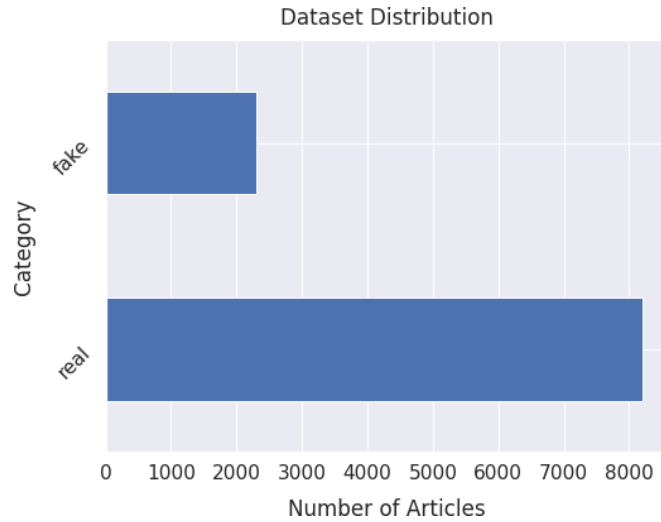


Figure-2: Dataset Distribution

The data collection consists of six columns and 10321 documents in total. After that, the data set is split into two more groups. Real and false are the categories. The quantity of papers, words, unique words, and most common terms are taken out of each category.

The class name is actual, there are 8043 documents, 45983 words, and 11578 unique terms. These are some selections of the most often used words. Words Most Often Used: ড়নহে 331, বাংািতেে 221, আটক 204, ঞেষয 201, ও 194, ড়বএনড়প 153, ২ 153, তসতেশ্বর 149, ড়ৌর 148, েধানমন্ত্ৰী 125.

The class name is false, there are 2278 documents, 16636 words, and 5184 unique terms. These are some selections of the most often used words. Words Most Often Used: দেড়নক 213, মড়েকণ্ঠ 213, এক 100, হতয় 74, তযভাতব 52, কারতে 50, সাতে 49, ১০টট 47, ড়েতয় 45, টাকা 43.

After data preparation, the total number of documents was 10321 and 186 data were cleaned and removed from the dataset. After examining the complete dataset, the most common word, number of titles, number of words, and number of unique words in each category are extracted. This process is repeated for each individual category among all the titles. In addition to counting the most frequent words, a table is created with the number of documents in each category, total number of words, number of unique words, and also shown by visualization below the table.

Table-3: Data Statistics

No	Total Documents	Total Words	Unique Words	Class Names
1	8043	45983	11578	Real
2	2278	16636	5184	Fake

Thus, by analyzing the real class, its total documents (8043), total words (45983), unique words (11578) and most frequent words have been extracted. By analyzing the fake class, its total documents (2278), total words (16636), unique words (5184) and most frequent words have been extracted. Below is the graph of dataset analysis:

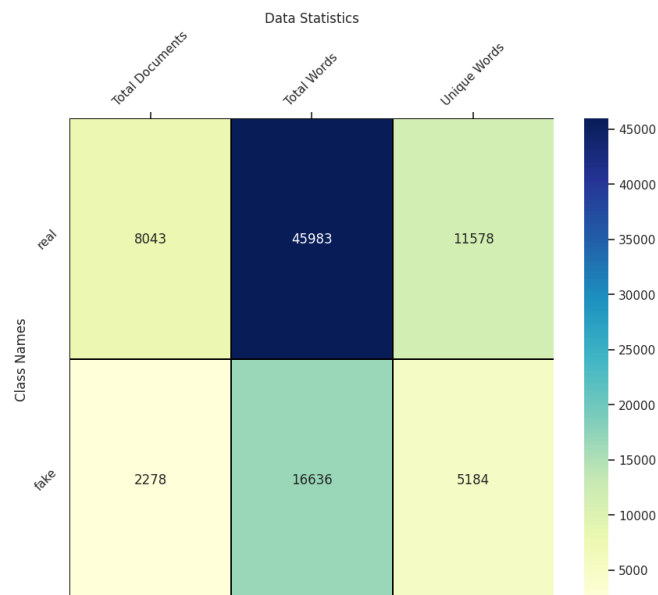


Figure-3: Dataset Statistics

The dataset's document length and frequency were measured following the visualization of the dataset. There are three minimum and six average lengths for the document, with a maximum length of 135 and a minimum of 3. The length frequency distribution graph is displayed below:

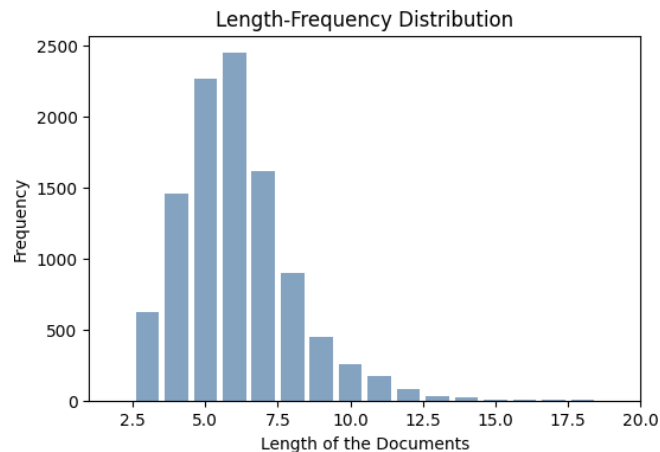


Figure-4: Length-frequency distribution

3.6. Label Encoding

Label encoding is used in the encoding of this dataset. In order to transform character type data into numeric index data that may be transformed into a machine-readable format, the process is known as level encoding. Level encoding is used to address the issue of computers not understanding categorical data properly or producing subpar outcomes when working with various machine algorithms. This is a crucial phase in supervised learning for data sets. In essence, categorical language is converted to numerical language using the "LabelEncoder()" method. Two categories have been created from our dataset, with level encoding applied to each.

Class Names: `[] ['fake' 'real']`.

Shape of Encoded Corpus `=====> (10321, 300)`.

The dataset is divided once all of the data has been converted to numerical form. Three sections make up the dataset: training, test, and validation data. 7430 of the 10321 total data are for training, while 1033 are for testing. There are 1033 pieces of data that are retained for validation checks.

3.7. Tokenizer

Tokenization is the process of dividing a text document into smaller sections, such as a phrase, paragraph, or whole text. When dealing with textual data, tokenization is frequently utilized. We have also tokenized our dataset. Our data collection had 10321

total documents, from which 14294 distinct tokens were extracted. We tokenize our data collection in two different ways. Specifically, padded and encoded sequences.

i. Encoded Sequences-

Encoded Sequences
মান্নার ১৪মে মেঘবাড়ষকী
[3786, 1, 2928]

ii. Padded Sequences-

Padded Sequences													
মান্নার ১৪মে মেঘবাড়ষকী													
[3786	1	2928	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0]

3.8. Recurrent Neural Network (RNN)

One type of neural network that is frequently utilized in natural language processing (NLP) is the recurrent neural network (RNN). This model also plays a useful role in modeling sequence data. That is, this model is more helpful in generating predictive results on sequential data i.e. recognizing the sequential characteristics of the data and using it to predict the next possible scenario. For example, Natural Language Processing (NLP), used in machine translation and speech recognition.

3.8.1. Validation and Training accuracy

Table-4: Validation and Training accuracy table for RNN

Epoch No	Validation Accuracy	Training Accuracy
Epoch 1	0.9150	0.8051
Epoch 2	0.9408	0.9474
Epoch 3	0.9429	0.9833

Epoch 4	0.9446	0.9948
Epoch 5	0.9408	0.9978
Epoch 6	0.9413	0.9995
Epoch 7	0.9408	0.9993
Epoch 8	0.9451	0.9997
Epoch 9	0.9403	0.9995
Epoch 10	0.9419	0.9996

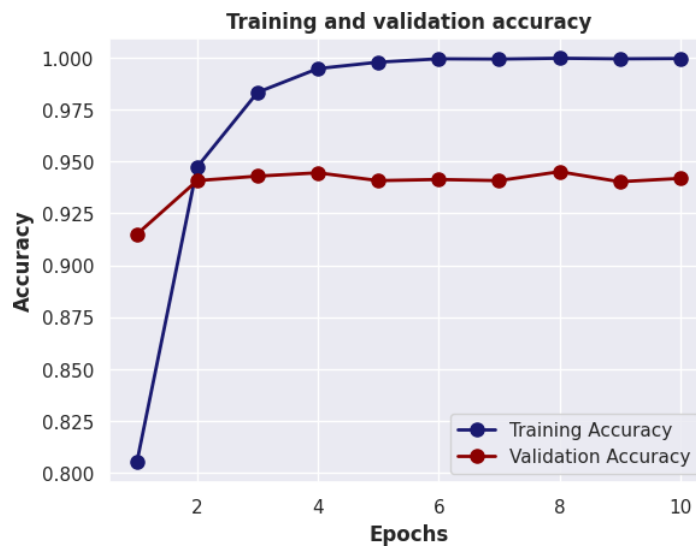


Figure-5: Epochs vs Training & Validation Accuracy Plot for RNN

The multiclass unbalanced classification issue is the reason why the validation accuracy has not increased by more than 95%, as can be seen from the accuracy plot. Furthermore, the performance of the model may be enhanced by appropriately fine-tuning the vocabulary sizes.

3.8.2. Validation and Training loss

Table-5: Validation and Training loss for RNN

No	Validation Loss	Training Loss
Epoch 1	0.2418	0.4450
Epoch 2	0.1752	0.1492
Epoch 3	0.1902	0.0559
Epoch 4	0.2203	0.0199
Epoch 5	0.2417	0.0084
Epoch 6	0.2664	0.0039

Epoch 7	0.2836	0.0030
Epoch 8	0.3101	0.0020
Epoch 9	0.3131	0.0019
Epoch 10	0.3256	0.0015

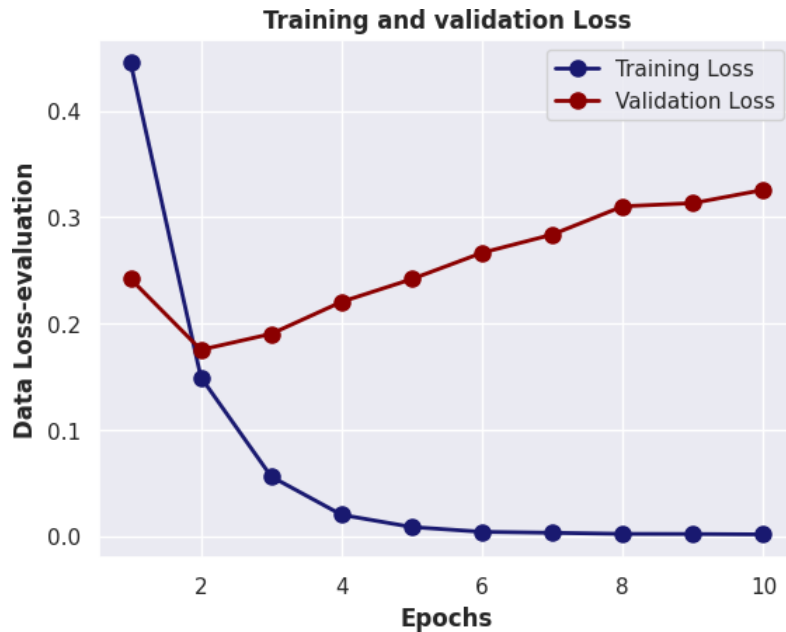


Figure-6: Epochs vs Training & Validation Loss Plot for RNN

3.8.3. Classification Report Model Performance

Table-6: Classification report of RNN Algorithm

	precision	recall	f1-score	support
Fake	88.94	84.86	86.85	218.000000
Real	96.00	97.18	96.59	815.000000
Accuracy	94.58	94.58	94.58	0.945789
Macro avg	92.47	91.02	91.72	1033.000000
Weighted avg	94.51	94.58	94.53	1033.000000

We can observe that all of the classes are fairly categorized by looking at the accuracy, recall, and f1-score.

3.8.4. Confusion matrix

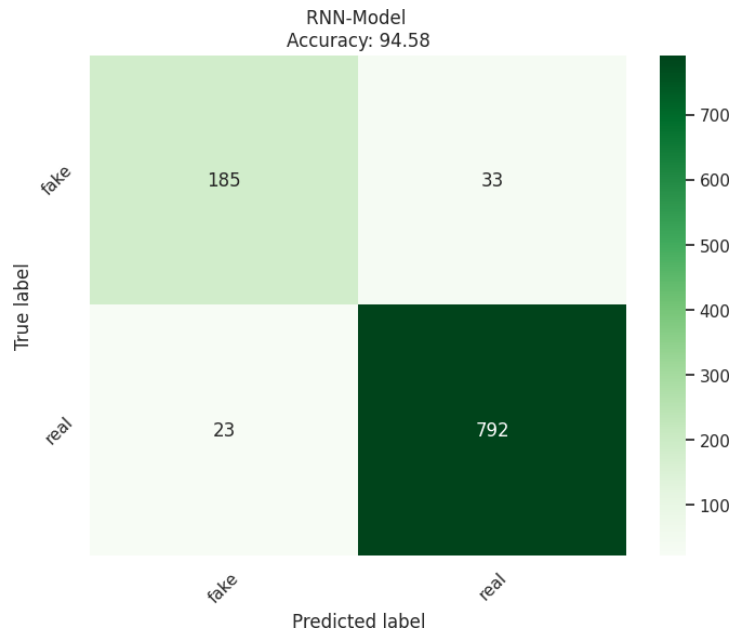


Figure-7: Confusion matrix for RNN

A clear picture of the number of documents successfully categorized in each class and the classes that experience confusion during classification is given by the confusion matrix. It is evident that a greater number of incorrectly categorized results are returned by the genuine category.

3.8.5. Testing RNN algorithm with our own news

Table-7: Test RNN algorithm using dummy news

Sample News	Class	Accuracy %
'অবতেতষ জানা তগি 'নাড়বি জাতনা' তপাস্টাতরর রহস্য!'	Fake	94.58

3.9. Long Short-Term Memory (LSTM)

Deep learning is what the long short-term memory network does. The brain network that makes information survive is hierarchical. LSTM is a special type of recurrent neural network used to solve vanishing gradient problems and sequence prediction problems. LSTM is designed to solve the problem of rnn and machine learning algorithms. LSTM cells effectively add long-term memory to learn more parametrically. LSTM models can

be converted to Python using Keras library. This model is used in complex problems like machine translation, speech recognition.

3.9.1. Validation and Training accuracy

Table-8: Validation and Training accuracy table for LSTM

Epoch No	Validation Accuracy	Training Accuracy
Epoch 1	0.8724	0.8287
Epoch 2	0.9295	0.9343
Epoch 3	0.9392	0.9682
Epoch 4	0.9397	0.9803
Epoch 5	0.9424	0.9894
Epoch 6	0.9462	0.9934
Epoch 7	0.9419	0.9948
Epoch 8	0.9370	0.9958
Epoch 9	0.9456	0.9969
Epoch 10	0.9403	0.9980

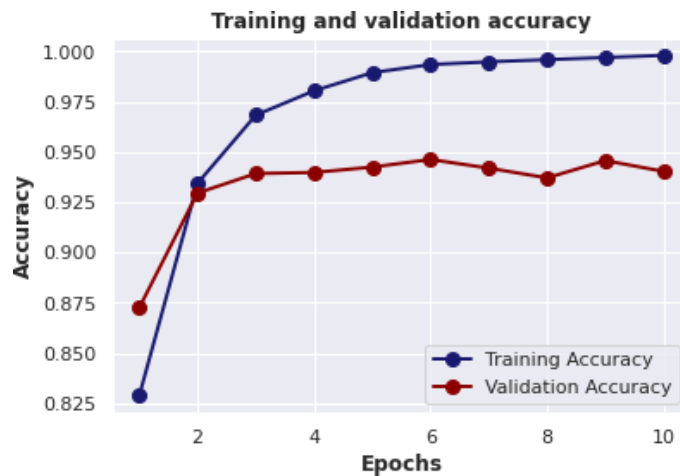


Figure-8: Epochs vs Training & Validation Accuracy Plot for LSTM

The multiclass unbalanced classification issue is the reason why the validation accuracy has not increased by more than 95%, as can be seen from the accuracy plot. Furthermore, the performance of the model may be enhanced by appropriately fine-tuning the vocabulary sizes.

3.9.2. Validation and Training loss

Table-9: Validation and Training loss for LSTM

No	Validation Loss	Training Loss
Epoch 1	0.3278	0.4054
Epoch 2	0.1802	0.1820
Epoch 3	0.1747	0.0943
Epoch 4	0.1896	0.0576
Epoch 5	0.2189	0.0397
Epoch 6	0.2443	0.0233
Epoch 7	0.2663	0.0183
Epoch 8	0.2854	0.0127
Epoch 9	0.3239	0.0098
Epoch 10	0.3729	0.0062

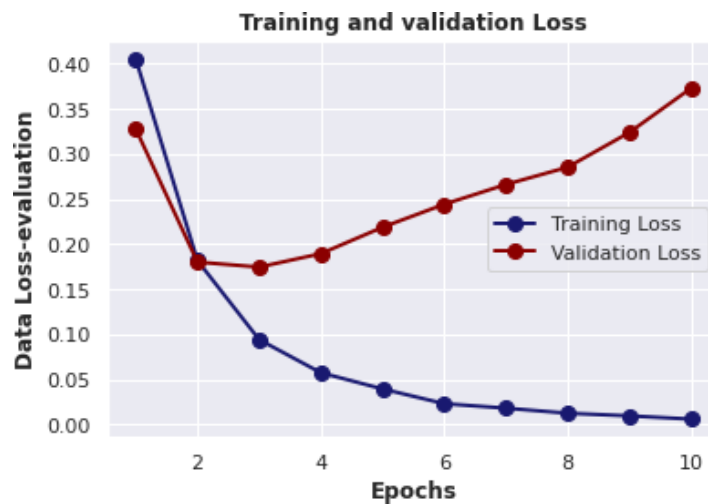


Figure-9: Epochs vs Training & Validation Loss Plot for LSTM

3.9.3. Classification Report Model Performance

Table-10: Classification report of LSTM Algorithm

	Precision	recall	f1-score	support
Fake	80.25	87.61	83.77	218.000000
Real	96.60	94.23	95.40	815.000000
Accuracy	92.84	92.84	92.84	0.928364
Macro avg	88.43	90.92	89.59	1033.000000
Weighted avg	93.15	92.84	92.95	1033.000000

We can observe that all of the classes are fairly categorized by looking at the accuracy, recall, and f1-score.

3.9.4. Confusion matrix

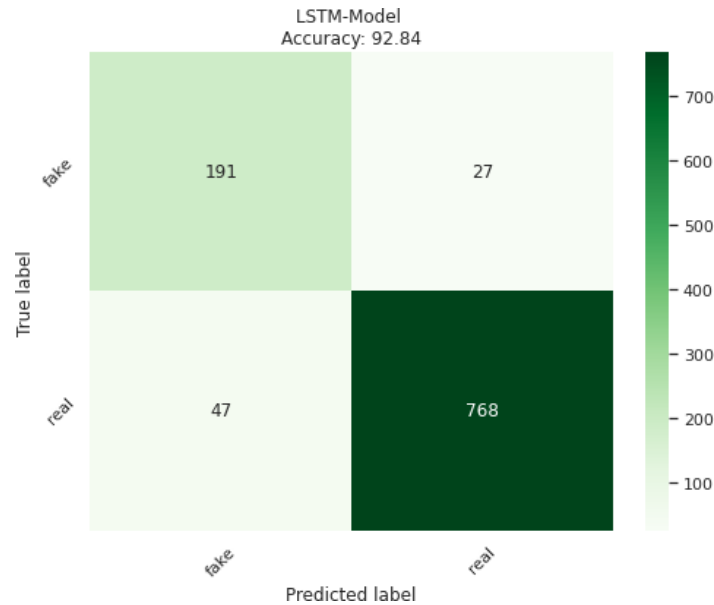


Figure-10: Confusion matrix for LSTM

A clear picture of the number of documents successfully categorized in each class and the classes that experience confusion during classification is given by the confusion matrix. It is evident that a greater number of incorrectly categorized results are returned by the genuine category.

3.9.5. Testing LSTM algorithm with our own news

Table-11: Test LSTM algorithm using dummy news

Sample News	Class	Accuracy %
'অবতেতষ জানা তগি 'নাড়বি জাতনা'তপাস্টাতরর রহসয!'	Fake	92.84

3.10. Bidirectional Long Short-Term Memory (Bi-LSTM)

BiLSTM, or Bidirectional LSTM, is a model for processing sequences. Two LSTM make up this sequence processing model: a) one without a backward input and b) one without a forward input. In NLP, this method is often utilized. Numerous NLP applications, including handwritten identification, protein structure prediction, and audio recognition,

employ this approach. able to show enhanced efficiency when solving sequential classification challenges.

3.10.1. Validation and Training accuracy

Table-12: Validation and Training accuracy table for Bi-LSTM

Epoch No	Validation Accuracy	Training Accuracy
Epoch 1	0.8994	0.8166
Epoch 2	0.9220	0.9427
Epoch 3	0.9429	0.9812
Epoch 4	0.9150	0.9929
Epoch 5	0.9419	0.9970
Epoch 6	0.9284	0.9977
Epoch 7	0.9424	0.9985
Epoch 8	0.9478	0.9991
Epoch 9	0.9386	0.9993
Epoch 10	0.9473	0.9995

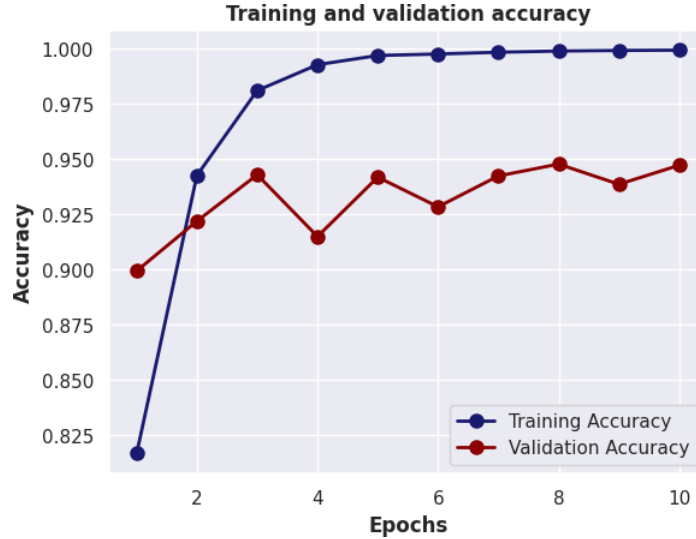


Figure-11: Epochs vs Training & Validation Accuracy Plot for Bi-LSTM

The multiclass unbalanced classification issue is the reason why the validation accuracy has not increased by more than 95%, as can be seen from the accuracy plot. Furthermore, the performance of the model may be enhanced by appropriately fine-tuning the vocabulary sizes.

3.10.2. Validation and Training loss

Table-13: Validation and Training loss for Bi-LSTM

No	Validation Loss	Training Loss
Epoch 1	0.2917	0.4428
Epoch 2	0.2003	0.1629
Epoch 3	0.2049	0.0625
Epoch 4	0.3262	0.0221
Epoch 5	0.2817	0.0119
Epoch 6	0.3112	0.0095
Epoch 7	0.3409	0.0056
Epoch 8	0.3273	0.0037
Epoch 9	0.4262	0.0022
Epoch 10	0.3513	0.0029

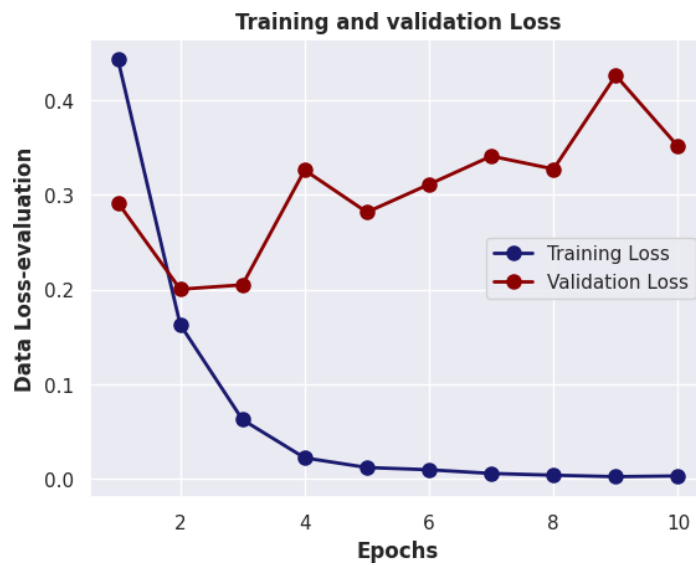


Figure-12: Epochs vs Training & Validation Loss Plot for Bi-LSTM

3.10.3. Classification Report Model Performance

Table-14: Classification report of Bi-LSTM Algorithm

	precision	recall	f1-score	support
Fake	86.98	85.78	86.37	218.000000

Real	96.21	96.56	96.39	815.000000
Accuracy	94.29	94.29	94.29	0.942885
Macro avg	91.59	91.17	91.38	1033.000000
Weighted avg	94.26	94.29	94.27	1033.000000

We can observe that all of the classes are fairly categorized by looking at the accuracy, recall, and f1-score.

3.10.4. Confusion matrix

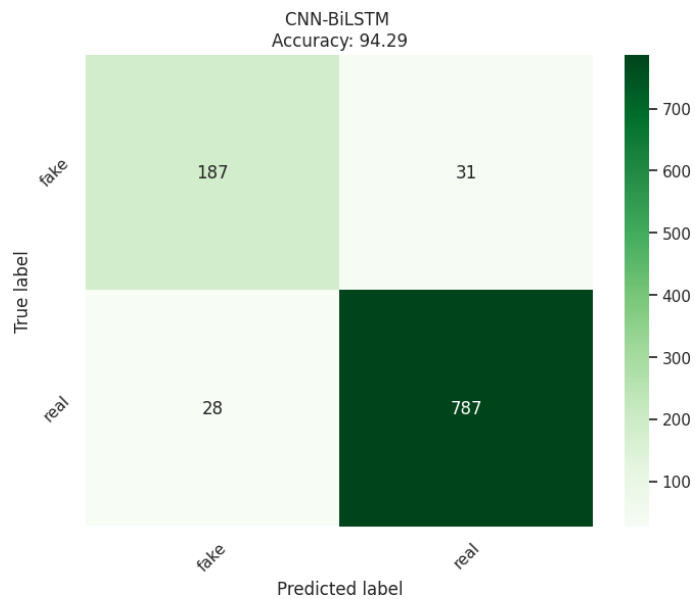


Figure-13: Confusion matrix for Bi-LSTM

A clear picture of the number of documents successfully categorized in each class and the classes that experience confusion during classification is given by the confusion matrix. It is evident that a greater number of incorrectly categorized results are returned by the genuine category.

3.10.5. Testing Bi-LSTM algorithm with our own news

Table-15: Test Bi-LSTM algorithm using dummy news

Sample News	Class	Accuracy %
'অবতেষ জানা তগি 'নাড়বি জাননা'তপাস্টাতরর রহসয!'	Fake	94.29

3.11. Gated recurrent units (GRU)

The GRU model, which was first presented by Kyunghyun Cho in 2014, is a gating mechanism system for recurrent neural networks. The GRU model essentially employs node sequencing to carry out memory and clustering-related machine learning tasks. It essentially belongs to the model of recurrent neural networks. While GRU and LSTM are comparable, GRU has one less parameter than LSTM. as GRU lacks an output gate. GRU's two gates are updated and reset. The three gates of an LSTM, on the other hand, are input, output, and forget. This indicates that since GRU has fewer gates than LSTM, it is less complicated than LSTM.

How GRU functions: A GRU cell functions much in the same way as an LSTM or RNN cell. It receives two inputs, hidden state and X_t At every timestamp t , take the H_{t-1} from the timestamp $t-1$ before it. Then, a fresh hidden state, H_t , is created and transferred to the next timestamp again.

3.11.1. Validation and Training accuracy

Table-16: Validation and Training accuracy table for GRU

Epoch No	Validation Accuracy	Training Accuracy
Epoch 1	0.9290	0.8357
Epoch 2	0.9424	0.9563
Epoch 3	0.9424	0.9801
Epoch 4	0.9429	0.9907
Epoch 5	0.9462	0.9946
Epoch 6	0.9467	0.9964
Epoch 7	0.9424	0.9978
Epoch 8	0.9440	0.9983
Epoch 9	0.9333	0.9987
Epoch 10	0.9370	0.9988

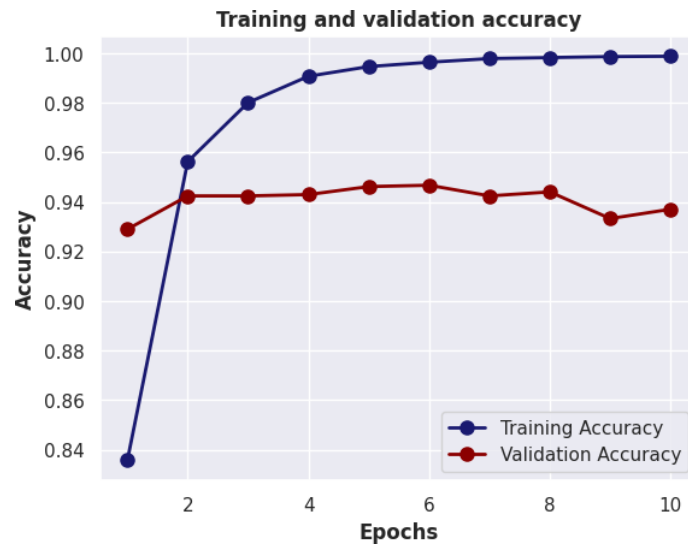


Figure-14: Epochs vs Training & Validation Accuracy Plot for GRU

The multiclass unbalanced classification issue is the reason why the validation accuracy has not increased by more than 95%, as can be seen from the accuracy plot. Furthermore, the performance of the model may be enhanced by appropriately fine-tuning the vocabulary sizes.

3.11.2. Validation and Training loss

Table-17: Validation and Training loss for GRU

No	Validation Loss	Training Loss
Epoch 1	0.1918	0.3934
Epoch 2	0.1663	0.1245
Epoch 3	0.1827	0.0603
Epoch 4	0.2199	0.0300
Epoch 5	0.2201	0.0176
Epoch 6	0.2769	0.0113
Epoch 7	0.3106	0.0065
Epoch 8	0.3180	0.0066
Epoch 9	0.3510	0.0049
Epoch 10	0.3614	0.0041

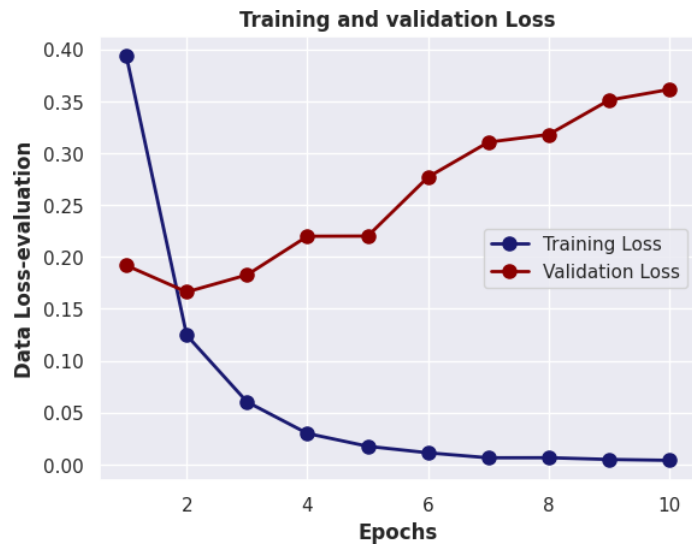


Figure-15: Epochs vs Training & Validation Loss Plot for GRU

3.11.3. Classification Report Model Performance

Table-18: Classification report of GRU Algorithm

	precision	recall	f1-score	support
Fake	82.17	86.70	84.38	218.000000
Real	96.39	94.97	95.67	815.000000
Accuracy	93.22	93.22	93.22	0.932236
Macro avg	89.28	90.83	90.02	1033.000000
Weighted avg	93.39	93.22	93.29	1033.000000

We can observe that all of the classes are fairly categorized by looking at the accuracy, recall, and f1-score.

3.11.4. Confusion matrix

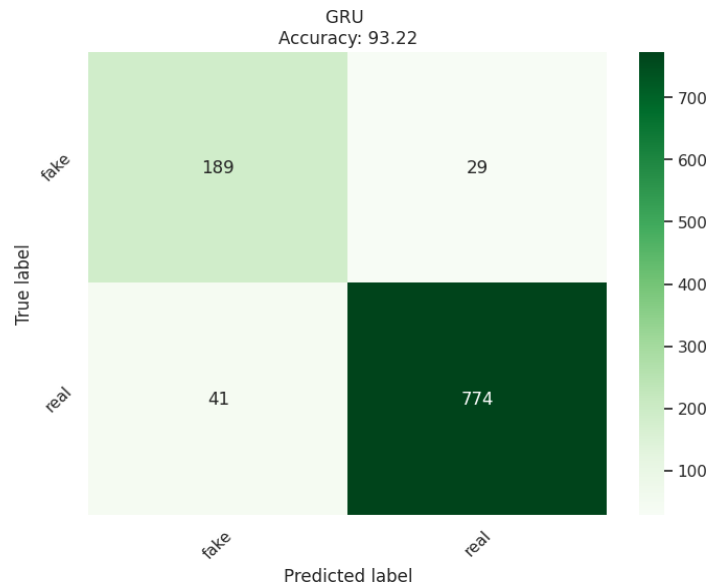


Figure-16: Confusion matrix for GRU

A clear picture of the number of documents successfully categorized in each class and the classes that experience confusion during classification is given by the confusion matrix. It is evident that a greater number of incorrectly categorized results are returned by the genuine category.

3.11.5. Testing GRU algorithm with our own news

Table-19: Test GRU algorithm using dummy news

Sample News	Class	Accuracy %
'অবতেষ জানা তগি 'নাড়বি জাতনা'তপাস্টাতরর রহসয!'	Fake	93.22

3.12. Bidirectional Encoder Representations from Transformers (BERT)

One deep learning approach is called Bidirectional Encoder Representations from Transformers (BERT); where each output is associated with each input, thereby predicting or calculating the dataset based on the association. This model is an open-source machine learning framework that is commonly used for natural language processing (NLP). It is intended to make the meaning of some of the text's cryptic terminology easier to grasp. Since Wikipedia content is used to pre-train the BERT framework, queries and responses may be derived from the dataset.

3.12.1. Dataset Description for BERT Algorithm

- Show bar graph authentic headlines length & content length top 100

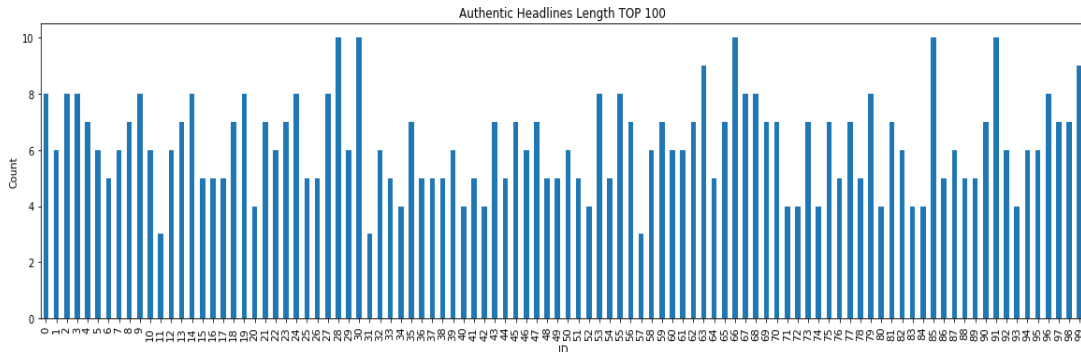


Figure-17: Authentic Headlines Length TOP 100

By describing the dataset in the BERT algorithm, it can be seen that the first 100 documents in the authentic headline length graph are taken. Where there are documents of all types of length. The length of the first 100 authentic headline documents is shown visually in a graph together.

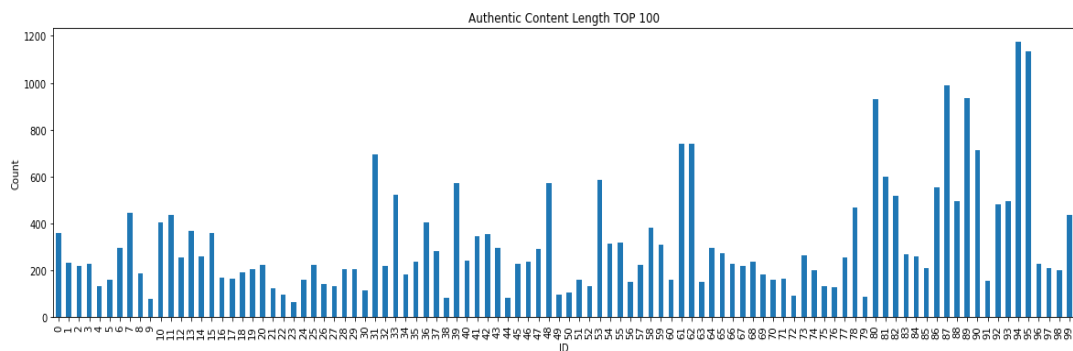


Figure-18: Authentic Content Length TOP 100

By describing the dataset in the BERT algorithm, it can be seen that the content length graph of the authentic headlines is made with the first 100 documents. Where there are documents of all types of content, small and large. The length of the content of the first 100 authentic headlines is visualized in a graph together.

- Show distance in authentic headlines length & content length top 100

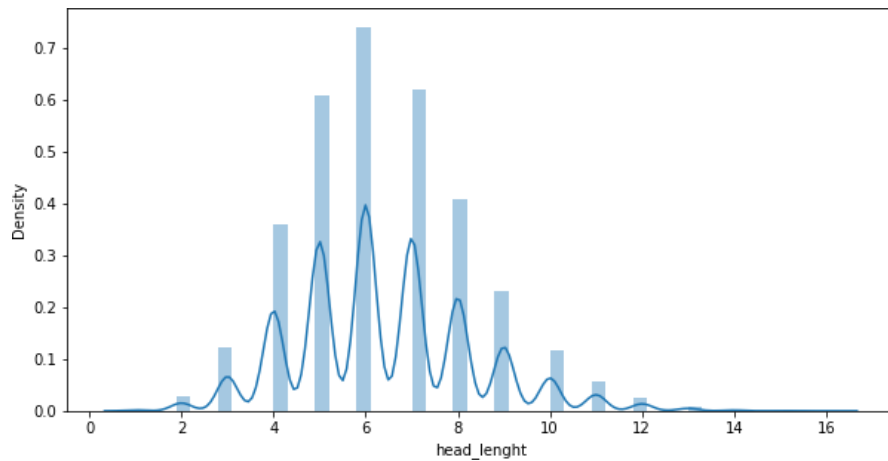


Figure-19: Authentic Headline Length TOP 100

By describing the dataset in the BERT algorithm, it can be seen that the distance graph of the authentic headline length is made with the first 100 documents. Where there are distances of all types of headline length, small and large. The distances of the first 100 authentic headline lengths are shown visually in the graph together.

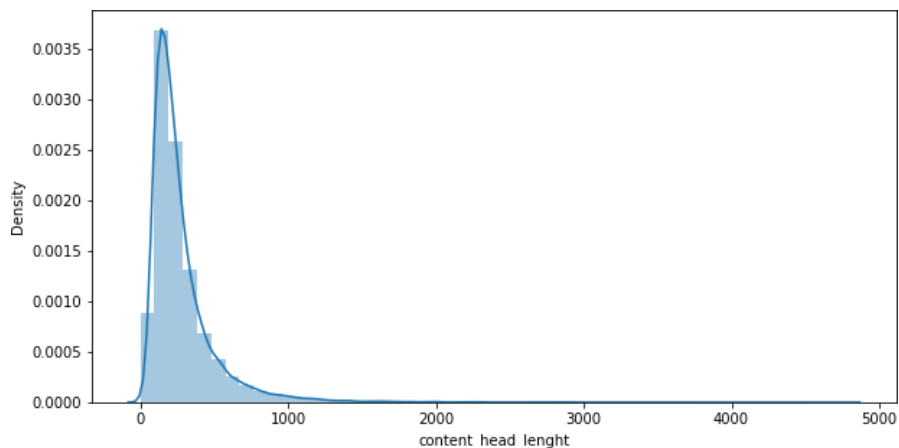


Figure-20: Authentic Content Headline Length TOP 100

By describing the dataset in the BERT algorithm, it can be seen that the authentic content headline length distance graph is constructed with the first 100 documents. Where there are all types of authentic content headline length distance. The distances of the first 100 authentic content headline lengths are shown visually in the graph together.

- Show bar graph fake headlines length & content length top 100

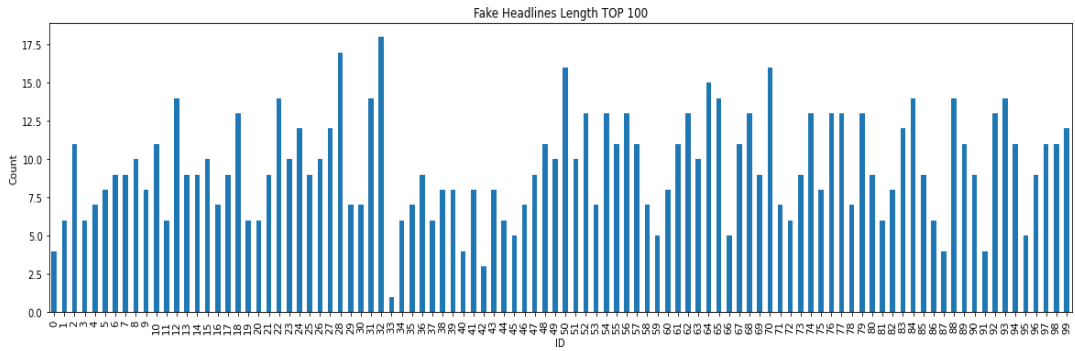


Figure-21: Fake Headlines Length TOP 100

By describing the dataset in the BERT algorithm, it can be seen that the first 100 documents in the fake headline length graph are taken. Where there are documents of all types of length. The length of the first 100 fake headline documents is shown visually in a graph together.

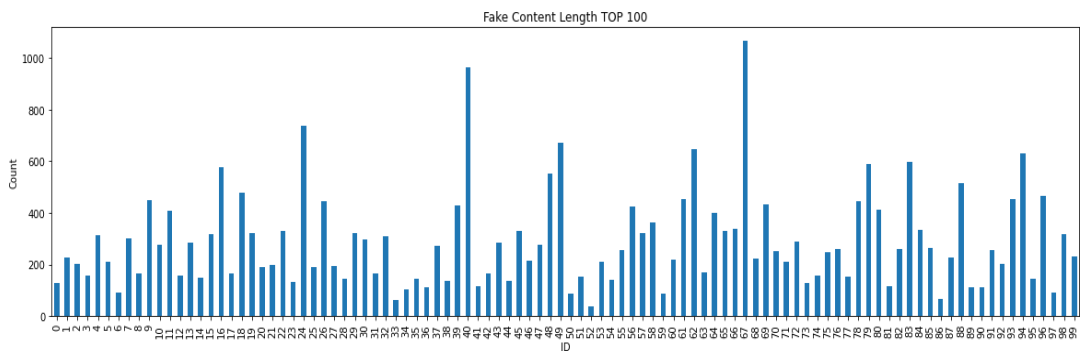


Figure-22: Fake Content Headline Length TOP 100

By describing the dataset in the BERT algorithm, it can be seen that the content length graph of the fake headlines is made with the first 100 documents. Where there are documents of all types of content, small and large. The length of the content of the first 100 fake headlines is visualized in a graph together.

- Show distance in fake headlines length & content length top 100

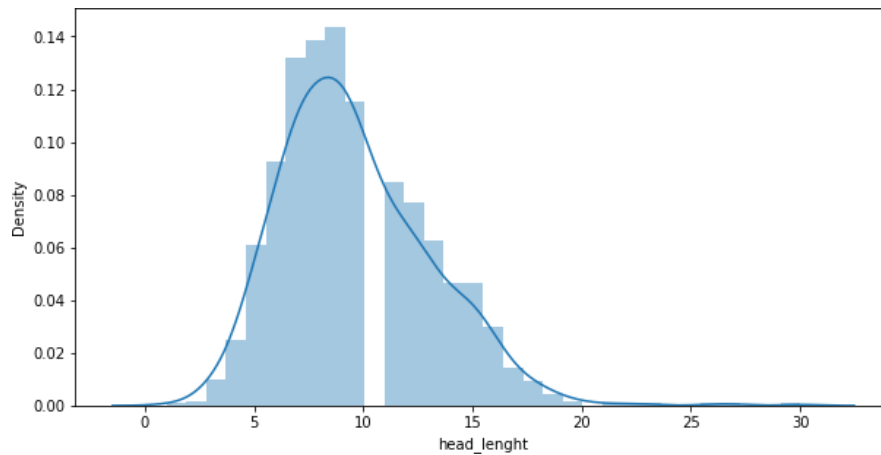


Figure-23: Fake Headline Length TOP 100

By describing the dataset in the BERT algorithm, it can be seen that the distance graph of the fake headline length is made with the first 100 documents. Where there are distances of all types of headline length, small and large. The distances of the first 100 fake headline lengths are shown visually in the graph together.

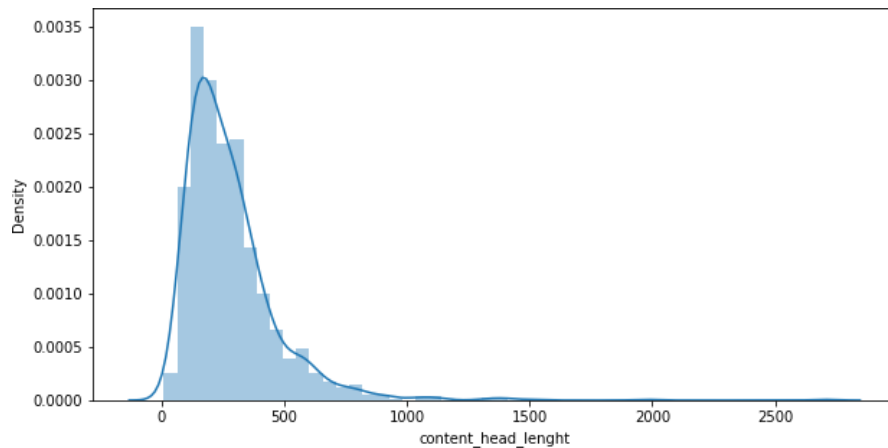


Figure-24: Fake Content Headline Length TOP 100

By describing the dataset in the BERT algorithm, it can be seen that the fake content headline length distance graph is constructed with the first 100 documents. Where there are all types of fake content headline length distance. The distances of the first 100 fake content headline lengths are shown visually in the graph together.

- Show word cloud image in authentic headlines



Figure-25: Word Cloud Image for Authentic Headlines

By describing the dataset to the BERT algorithm, it can be seen that the word cloud image of the authentic headline is shown. Here basically an image is created with a frame with the same letters of all the authentic headlines.

- Show word cloud image in fake headlines



Figure-26: Word Cloud Image for Fake Headlines

By describing the dataset to the BERT algorithm, it can be seen that the word cloud image of the fake headline is shown. Here basically an image is created with a frame with the same letters of all the fake headlines.

- Show stop words in authentic headlines

(না, 'ও', 'শুরু', 'হতব', 'কতর', 'ডনতয়', 'জনয', 'তেতক', 'েই', 'নেষন', 'সতে', 'তকাটট', 'করা', 'করতে',
'হতে', 'ডেতে', 'পর', 'েমে', 'কাজ', 'হাজার')
(57, 34, 23, 23, 23, 22, 21, 21, 19, 17, 17, 16, 16, 14, 14, 11, 11, 11, 11, 11)

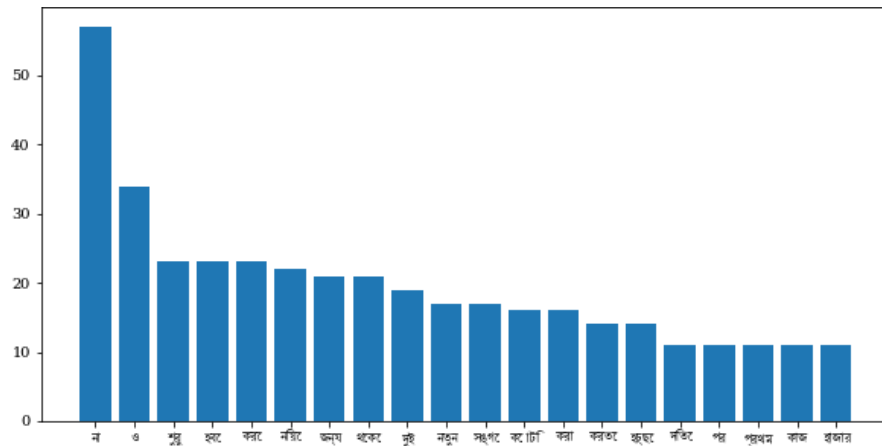


Figure-27: Stop Words for Authentic Headlines

By describing the dataset to the BERT algorithm, it can be seen that the most used words in authentic headlines are shown and the number of times each word is also shown. Also, the most used words are visually displayed in the graph.

- Show stop words in fake headlines

(না, 'তয', 'কতর', 'ও', 'তেতক', 'ডনতয়', 'এই', 'জনয', 'ডেতিন', 'করতিন', 'এবার', 'করতে', 'ডক', 'করা', 'তবড়ে', 'হতব', 'েই', 'ষা', 'পর', 'আমার')
(101, 58, 53, 40, 36, 36, 36, 28, 26, 25, 25, 25, 25, 22, 20, 20, 20, 18, 17, 17)

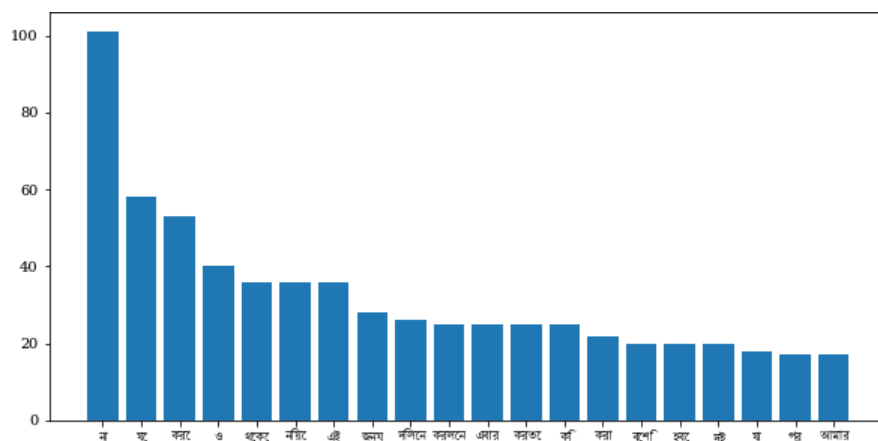


Figure-28: Stop Words for Fake Headlines

By describing the dataset to the BERT algorithm, it can be seen that the most used words in fake headlines are shown and the number of times each word is also shown. Also, the most used words are visually displayed in the graph.

- Top 10 most frequent words in authentic headlines

('১৯', 'তসত্তেধর', '২০১৮', 'েধানমন্ত্রী', 'পাড়কস্তান', 'িাখ', 'ভারে', 'আটক', 'বাংিাতেে', 'ডোর')
(56, 52, 51, 45, 33, 25, 23, 21, 20, 20)

By describing the dataset to the BERT algorithm, it can be seen that the top 10 most used words in authentic headlines are shown and the frequency of each word is also shown. Also, the words are shown sequentially i.e. from higher to lower numbers.

- Top 10 most frequent words in fake headlines

('দেড়নক', 'মডেকঠ', 'Bengal', 'Beats', 'এক', 'হতয়', 'সাতে', 'ডেতয়', 'কারতে', 'তেখ')
(151, 151, 139, 139, 41, 29, 24, 21, 20, 19)

By describing the dataset to the BERT algorithm, it can be seen that the top 10 most used words in authentic headlines are shown and the frequency of each word is also shown. Also, the words are shown sequentially i.e. from higher to lower numbers.

3.12.2. Training and validation loss

Table-20: Validation loss, Training loss and Accuracy for BERT

No	Validation Loss	Training Loss	Accuracy
Epoch 1	0.209000	0.241535	93.7951
Epoch 2	0.159097	0.144717	95.3824
Epoch 3	0.184832	0.077036	95.5748
Epoch 4	0.188217	0.067440	95.7191
Epoch 5	0.186699	0.062669	95.6710

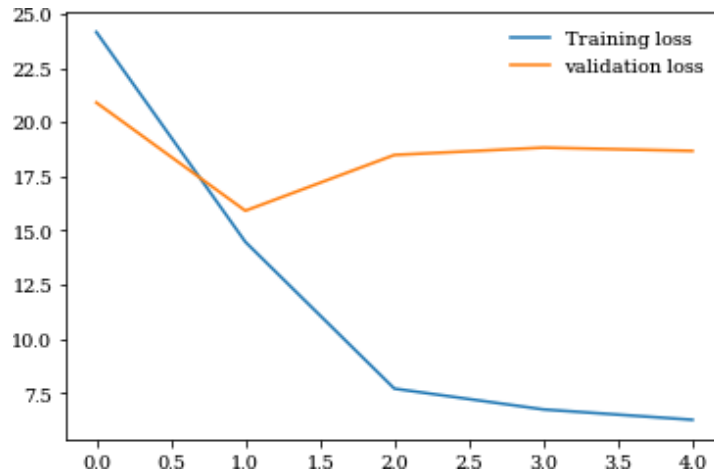


Figure-29: Epochs vs Training & Validation Loss Plot for BERT

3.12.3. Classification Report Model Performance

Table-21: Classification report of BERT algorithm

	precision	recall	f1-score	support
Fake (0)	0.83	0.80	0.81	260
Authentic (1)	0.97	0.98	0.97	1819
Accuracy			0.95	2079
Macro avg	0.90	0.89	0.89	2079
Weighted avg	0.95	0.95	0.95	2079

We can observe that all of the classes are fairly categorized by looking at the accuracy, recall, and f1-score.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Discussion

In the modern world, news is becoming more and more abundant. As the days pass, news stories on a variety of subjects are being produced. In addition to newspapers, news may also be found on a variety of news-based websites, social media platforms, blogs, and online news portals. A certain amount of writing for a newspaper cannot be repeated. However, there's no cap on how much you may write for websites, blogs, social media, or online news portals. As a result, the web news is published in a very thorough and evocative way. Consequently, there is an exponential growth in the quantity of news and this trend is continuing. If this growing news gets into the hands of some fraudsters and they misuse it, it will cause huge loss to the online news portals. Because this existing data/news sphere of online cannot be used in any way in the future. Besides, if the offline news is also harmed in the above way, then the people of the country will have a bad effect on such activities of the country and the people of the foreign countries will have a bad attitude about Bangladesh. But once a news article is published, we cannot understand or know whether it is authentic news or fake news. To detect this, digitized science, artificial intelligence, science has given different methods, such as machine learning, deep learning etc. with the help of different algorithms, we can detect which is real news and which is fake news. In this paper, some algorithms of deep learning are applied for detection of fake news. Applied algorithms are RNN, LSTM, BiLSTM, GRU and BERT.

Table-22: Classifiers Description

Classifier	Description
RNN	Through RNN, a class creates a cycle between its own nodes, in which the output of one node is used as the input of the next node.
LSTM	A subset of RNNs called LSTMs is able to identify long-term dependencies in sequence prediction issues.
Bi-LSTM	To provide an output that is more accurate, BiLSTM integrates input sequences with data from the past and the future.

GRU	To sort nodes, GRUs are usually used with machine learning's in-memory clustering and clustering techniques. Update and reset gates are the two gates that are used.
BERT	An open-source NLP framework is called BERT. BERT's purpose is to interpret unclear words by analyzing the surrounding content.

4.2. Experimental Results and Analysis

4 deep learning models in the paper namely RNN, LSTM, BiLSTM and GRU; And a machine learning model used in the paper called BERT. RNN, LSTM, BiLSTM and GRU deep learning models have 94.58% accuracy in RNN, 92.84% accuracy in LSTM, 94.29% accuracy in BiLSTM and 93.22% accuracy in GRU. Also, the machine learning model BERT has 95% accuracy. It can be seen that RNN, BiLSTM model using deep learning obtained the highest accuracy of 94.58%, 84.29% respectively and BERT model using machine learning obtained 95% accuracy. Here it can be seen that the highest accuracy is obtained using the BERT model of machine learning.

Table-23: Classifiers accuracy, recall and precision

Algorithm techniques	Accuracy %	Recall %	Precision %
RNN	94.58	94.58	94.58
LSTM	92.84	92.84	92.84
Bi-LSTM	94.29	94.29	94.29
GRU	93.22	93.22	93.22
BERT	95	95	95

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1. Impact on Society

The internet has given modern living a new dimension through virtual entertainment. In rural areas, people no longer read newspapers; instead, they use cellphones to browse and read news websites. All of it is accessible through virtual entertainment platforms such as YouTube, Messenger, Instagram, LinkedIn, Pinterest, Trembler, Snapchat, Viber, WhatsApp, Facebook, Twitter, Google, and others. Virtual entertainment timelines and news feeds are becoming overflowing with both important and irrelevant news. Social media is used by 70% of internet users globally. In children, the percentage is about 90%. According to a survey, Facebook is used by almost 80% of Bangladeshi online users. The extent of social communication has increased due to the use of the internet compared to earlier times. Individuals exchange information, thoughts, images, videos, and other media through various data and document formats thanks to technological advancement. Since none of these kinds of information and documents are distinctive or authentic, searching for a certain topic will provide poor results and no authentic news. This is a significant disadvantage. Data categorization is finished in order to solve this issue. That is, it will be very helpful in finding news and identifying fake news, providing that the news is divided into two groups, such as true and fake news. Assuming, that is, that you may use any type of real term to compose a search query for news. Therefore, it is necessary to first categorize news into two groups: actual and fraudulent, in order to recognize any sort of genuine information. It will be simpler and more credible for individuals from diverse backgrounds to read authentic news on the internet or through virtual entertainment by identifying false news if it is divided into two categories.

5.2. Impact on Environment

Since ancient times, information has piqued human curiosity. News from newspapers was read by humans since the beginning of time. On the other hand, as innovation advances, fewer people read newspapers for news. In the modern world, social media and the internet provide access to a wide variety of reliable information. We have instant

access to any sort of reliable information on events occurring anywhere in the world. We receive genuine climatic reports via the Meteorological Department online in a timely manner. We are able to get a wide range of reliable information over the internet, such as the locations of cyclones, floods, and twisters as well as global environmental conditions. We are able to quickly ascertain the current weather conditions in any nation or location. We don't base our decisions on what should happen sooner or later on the climate or weather forecast. In this manner it is possible to make causes out of elements of much greater accident or misfortune. So, any kind of authentic information contributes a lot.

5.3. Ethical Aspects

Every news outlet has a distinct publication approach. The media can adopt strategies on its own. Newspapers on social media typically receive bogus news from the media. For example, providing false information online about any individual related to financial, political, income tax, or other matters is dishonest. In any event, bogus news about persons is disseminated on social media in numerous places. Despite this, there is less false information available online now than there ever was. That suggests that everyone should practice morality and behave decently when using social media or the internet.

5.4. Sustainability

We must separate the actual data or report into distinct categories in order to maintain my data collecting in those categories. Thus, genuine news or reports fall into several types. From this point on, the unidentified data may be utilized for any worthwhile purpose, including the detection of bogus news. For this reason, a lengthy system is needed, whereby any kind of data may be preprocessed and split into two classes—real and fake—using various techniques. That's why long-term planning requires possessing an aircraft.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLEMENTATION FOR FUTURE RESEARCH

6.1. Summary of the Study

However, no news or data will be valuable until it can be transformed into actual news and categorized. Thus, the requirement for classification through the transformation of news into authentic news is growing daily. Data science algorithms may be categorized using several techniques for identifying false news, including deep learning and machine learning algorithms. They are used to classify news with fake news detection so that people can convert it to real news by looking at the headlines to understand what kind of news it is and whether it is authentic. Our research paper is divided into 2 categories i.e., real and fake. BLSTM, GRU, Uni-Gram, Machine Learning (LR, Multinomial NB, RF, SVM) have been used to classify these categories. 4 deep learning models in the paper namely RNN, LSTM, BiLSTM and GRU; And a machine learning model used in the paper called BERT. RNN, LSTM, BiLSTM and GRU deep learning models have 94.58% accuracy in RNN, 92.84% accuracy in LSTM, 94.29% accuracy in BiLSTM and 93.22% accuracy in GRU. Also, the machine learning model BERT has 95% accuracy. In Bangladesh, the number of e-news is growing daily. These reliable and categorized e-news may be found on many different websites, but viewers tend to favor those that offer breaking news and updated content often. Prothom Alo, Bangladesh Pratidin, Nayadiganta, Jugantor, Samakal, and other well-known websites are among the five that frequently offer reliable and confidential news. These websites are highly well-liked by users since their news is often reliable and discreet.

6.2. Conclusions

Fake news identification is not categorizing the news at the same rate that the amount of e-news being produced worldwide is increasing. Thus, not every news is applicable. All news can be used if a significant portion of them can be identified as false news using fake news detection. By using algorithms like the machine learning, deep learning, and other learning algorithms employed in this study article, data science may be categorized in several ways to detect false news. Real and fraudulent news and information are

separated into two groups. 4 deep learning models in the paper namely RNN, LSTM, BiLSTM and GRU; And a machine learning model used in the paper called BERT. RNN, LSTM, BiLSTM and GRU deep learning models have 94.58% accuracy in RNN, 92.84% accuracy in LSTM, 94.29% accuracy in BiLSTM and 93.22% accuracy in GRU. Also, the machine learning model BERT has 95% accuracy. It can be seen that RNN, BiLSTM model using deep learning obtained the highest accuracy of 94.58%, 84.29% respectively and BERT model using machine learning obtained 95% accuracy.

6.3. Implication for Further Study

In this study, we demonstrate how to use false news detection to categories vast volumes of material into several categories. All news might potentially be abused in the future if false news detection is able to classify such a vast number of news items. Fake news detection may be implemented using various data science techniques, including machine learning and deep learning algorithms, to categories different sorts of algorithms. More sophisticated algorithms will be employed in the future to categories the data by identifying false news. This study demonstrates the usage of only five algorithms; however, other algorithms will be included in the future to further classify such data with the identification of false news. Additionally, only news or data is separated into two categories in this paper: authentic and fraudulent. Additional categories will be added later. In the future, news headlines can be anticipated with more accuracy thanks to the algorithms utilized in this article.

REFERENCE

- [1] Kingaonkar, S. *et al.* (2023) 'Fake News Detection using Machine Learning', *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 07(03). doi:10.55041/ijsrem18181.
- [2] Gaikwad, T. *et al.* (2022) *Detection of fake news using Machine Learning Algorithms*, *International Journal for Research in Applied Science & Engineering Technology*, 10(XII).
- [3] Adedoyin, F. and Mariyappan, B. (2022) *Fake news detection using machine learning algorithms and recurrent neural networks* [Preprint]. doi:10.31124/advance.20751379.v1.
- [4] Choudhary, M. *et al.* (2021) 'A review of fake news detection methods using machine learning', *2021 2nd International Conference for Emerging Technology (INCET)* [Preprint]. doi:10.1109/incet51464.2021.9456299.
- [5] G. Senthilkumar and D. Ashok Kumar (2023) 'A comparative study on various machine learning algorithms for the prediction of fake news detections using bring feed new data sets', *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 9(1), pp. 131–142. doi:10.32628/cseit228691.
- [6] Rahman, M. *et al.* (2022) 'Political fake news detection from different news source on social media using Machine Learning Techniques', *AIUB Journal of Science and Engineering (AJSE)*, 21(2), pp. 110–117. doi:10.53799/ajse.v21i1.383.
- [7] Kulkarni, P. *et al.* (2021) 'Fake news detection using machine learning', *ITM Web of Conferences*, 40, p. 03003. doi:10.1051/itmconf/20214003003.
- [8] Choudhury, D. and Acharjee, T. (2022) 'A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers', *Multimedia Tools and Applications*, 82(6), pp. 9029–9045. doi:10.1007/s11042-022-12788-1.
- [9] GOMATHY, Dr.C.K. *et al.* (2021) 'THE FAKE NEWS DETECTION USING MACHINE LEARNING ALGORITHMS', *International Research Journal of Engineering and Technology*, 08(10), pp. 37–40.
- [10] Murugesan, S. and Pachamuthu, K. (2022) 'Fake news detection in the medical field using machine learning techniques', *International Journal of Safety and Security Engineering*, 12(6), pp. 723–727. doi:10.18280/ijss.120608.
- [11] Khanam, Z. *et al.* (2021) 'Fake news detection using machine learning approaches', *IOP Conference Series: Materials Science and Engineering*, 1099(1), p. 012040. doi:10.1088/1757-899x/1099/1/012040.
- [12] Kushwaha, N.S. and Singh, P. (2022) 'Fake news detection using machine learning: A comprehensive analysis', *Journal of Management and Service Science (JMSS)*, 2(1), pp. 1–15. doi:10.54060/jmss/002.01.001.
- [13] Ngada, O. and Haskins, B. (2020) 'Fake news detection using content-based features and machine learning', *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)* [Preprint]. doi:10.1109/csde50874.2020.9411638.
- [14] Lakshmanarao, A., Swathi, Y. and Kiran, Dr.T. (2019) 'An effecient fake news detection system using machine learning', *International Journal of Innovative Technology and Exploring Engineering*, 8(10), pp. 3125–3129. doi:10.35940/ijitee.j9453.0881019.
- [15] Lasotte, Y.B. *et al.* (2022) 'An Ensemble Machine Learning Approach for Fake News Detection and Classification Using a Soft Voting Classifier', *European Journal of Electrical Engineering and Computer Science*, 6(2). doi://dx.doi.org/10.24018/ejece.2021.6.2.409.

- [16] Sultana, R. *et al.* (2022) ‘An effective fake news detection on social media and online news portal by using Machine Learning’, *Australian Journal of Engineering and Innovative Technology*, pp. 95–106. doi:10.34104/ajeit.022.0950106.
- [17] K. Nath, P. Soni, Anjum, A. Ahuja, and R. Katarya, “Study of Fake News Detection using Machine Learning and Deep Learning Classification Methods,” *2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology, IEEE Xplore*, Aug. 01, 2021. <https://ieeexplore.ieee.org/document/9573583> (accessed Sep. 18, 2022).
- [18] Md. Y. Tohabar, N. Nasrah, and A. M. Samir, “Bengali Fake News Detection Using Machine Learning and Effectiveness of Sentiment as a Feature,” *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, Aug. 2021, doi: <https://doi.org/10.1109/icievicivpr52578.2021.9564138>.
- [19] S. Pandey, S. Prabhakaran, N. V. Subba Reddy, and D. Acharya, “Fake News Detection from Online media using Machine learning Classifiers,” *Journal of Physics: Conference Series*, vol. 2161, no. 1, p. 012027, Jan. 2022, doi: <https://doi.org/10.1088/1742-6596/2161/1/012027>.
- [20] A. Santhosh Kumar, P. Kalpana, K. Sharan Athithya, and V. Sri Ajay Sundar, “Fake News Detection on Social Media Using Machine Learning,” *Journal of Physics: Conference Series*, vol. 1916, no. 1, p. 012235, May 2021, doi: <https://doi.org/10.1088/1742-6596/1916/1/012235>.

Shihab

ORIGINALITY REPORT

20%

SIMILARITY INDEX

16%

INTERNET SOURCES

10%

PUBLICATIONS

4%

STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	Submitted to University of North Carolina, Greensboro Student Paper	2%
3	Submitted to Liverpool John Moores University Student Paper	1%
4	www.researchgate.net Internet Source	1%
5	"Inventive Communication and Computational Technologies", Springer Science and Business Media LLC, 2023 Publication	1%
6	Sheikh Sadi Bandan, Sabid Ahmed Sunve, Shaklian Mostak Romel. "A Deep Learning Approach for Bengali News Headline Categorization", 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023 Publication	1%