

**BANGLA FAKE NEWS IDENTIFICATION USING DEEP LEARNING AND
MACHINE LEARNING**

BY

**MD. Shakhawat Hossain
ID: 201-15-3135**

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

Supervised By

Faria Nishat Khan
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mohammad Asifur Rahim
Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2024

APPROVAL

This Project titled “**Bangla Fake News Identification Using Deep Learning and Machine Learning**”, submitted by **MD. Shakhawat Hossain, ID:201-15-3135** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partially required fulfillment for B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **26.01.2024**

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori (SRH)
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University Chairman

Internal Examiner

Nazmun Nessa Moon (NNM)
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dewan Mamun Raza (DMR)
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Md. Arshad Ali (DAA)
Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science & Technology University

DECLARATION

I hereby announce that I completed this project under the guidance of **Faria Nishat Khan**, **Lecturer, Department of CSE** Daffodil International University. I further announce that neither this project nor any part of it has been forwarded to any other institution for the purpose of receiving a degree or diploma.

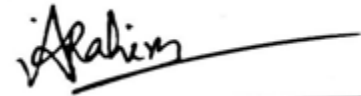
Supervised by:



Faria Nishat Khan

Lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:



Mohammad Asifur Rahim

Lecturer
Department of Computer Science and Engineering
Daffodil International University

Submitted by:



MD. Shakhawat Hossain

ID:201-15-3135
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

First, I express our heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible for me to complete the final year project/internship successfully. The real spirit of goal achievement lies in the pursuit of perfection and severe discipline.

I'd like to share my heartfelt gratitude to my supervisor, **Faria Nishat Khan** (Lecturer) supervisor, and **Mohammad Asifur Rahman** (Lecturer) co-supervisor, BSc in CSE Program, Daffodil International University. Her never-ending diligence, scholarly instruction, relentless motivation, continuous and diligent supervision, helpful critique, insightful suggestions, and reading many imperfect drafts and improving them at any point enabled the project to be completed. I am especially grateful to my honorable teachers.

I'd like to share my heartfelt gratitude to the Head, Professor **Dr. Sheak Rashed Haider Noori** Department of CSE, for his generous assistance in completing my thesis, and I'd like to express my gratitude to the Daffodil International University (DIU) staff for allowing me access to every type of equipment and archive resources to acquire information and clarify my understandings. I must express my gratitude for the advice provided by other supervisors and lecturers, who have assisted me in clarifying my understanding as well as instilling in me the value of properly executing the project report while maintaining good information and consistency.

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

ABSTRACT

Identification Of Bangla Fake News is very challenging, mostly when there is so much news from multiple online sources like social media, online news, and other online platforms. Detecting fake news is comparable in significance to eliminating crime from society. In this era of digital communication, the ease of spreading violence and manipulating public perception has reached unprecedented levels. Genuine news is frequently distorted into fake narratives for personal motives. In our daily lives, where we extensively utilize online platforms for various purposes, the relentless bombardment of fake news has become pervasive. To safeguard our well-being, it is imperative to put an end to this toxicity, and the key to achieving this lies in the detection of fake news.

My research is centered on addressing this crucial aspect. I endeavored to construct a framework for detecting false news, utilizing a significant dataset consisting of Bengali news information sourced from diverse online outlets. Employing both approaches in machine learning and as well as deep learning, this study revealed notable outcomes. In this study we used models such as Bidirectional GRU Model, Convolutional Neural Network (CNN) + Long Short-Term Memory (LSTM) + GlobalMaxPooling1D layers (Hybrid Model), Long Short-Term Memory (LSTM), and 1D Convolutional Neural Network (CNN). The Bidirectional GRU Model exhibited the highest accuracy of 99.13%. Despite facing limitations due to the performance constraints of our device, our research lays the foundation for developing an application dedicated to distinguishing between fake and authentic news.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgements	iii
Abstract	iv
Chapters	
Chapter 1: Introduction	1-5
1.1 Introduction	1-2
1.2 Motivation	2-3
1.3 Rationale of the Study	3-4
1.4 Research Questions	4
1.5 Expected Output	4-5
1.6 Finance and Project Management	5
1.7 Report Layout	5
Chapter 2: Background	6-11
2.1 Preliminaries/Terminologies	6
2.2 Related Works	6-10
2.3 Summary	10
2.4 Scope of the Problem	10-11

2.5 Challenges	11
Chapter 3: Research Methodology	12-29
3.1 Instrumentation and Research Subject	12
3.2 Data Collection Procedure	12-16
3.3 Statistical Analysis	16-20
3.4 Proposed Methodology/Applied Mechanism	20-28
3.5 Implementation Requirements	28-29
Chapter 4: Discussion and Experimental Results	30-32
4.1 Experimental Setup	30
4.2 Experimental Results & Analysis	30-31
4.3 Discussion	32
Chapter 5: Environment, Impact on Society and Sustainability	33-34
5.1 Impact on Society	33
5.2 Impact on Environment	33-34
5.3 Ethical Aspects	34
5.4 Sustainability Plan	34
Chapter 6: Summary, Conclusion, Recommendation And Implication for Future Research	35-34
6.1 Summary of the Study	35
6.2 Conclusions	35-36
6.3 Implication for Further Study	36

References	37-39
Plagiarism report	40

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.2.1: Ratio of authentic data and fake data	13
Figure 3.2.2: Data Collection Workflow	14
Figure 3.3.1: Graphical Representation of Train Data	17
Figure 3.3.2: Graphical Representation of Test Data	18
Figure 3.3.3: Train Plot NLP Image	19
Figure 3.3.4: Test Plot NLP Image	20
Figure 3.4.1: Confusion Matrix	21
Figure 3.4.2: ROC Curve	22
Figure 3.4.3: Confusion Matrix	23
Figure 3.4.4: ROC Curve	24
Figure 3.4.5: Confusion Matrix	25
Figure 3.4.6: ROC Curve	26
Figure 3.4.7: Confusion Matrix	27
Figure 3.4.8: ROC Curve	28

LIST OF TABLES

TABLES	PAGE NO
Table 2.2.1: Comparative Analysis	8-10
Table 3.2.1: Sample of Authentic Data	15
Table 3.2.2: Sample of Fake Data	15
Table 3.2.3: Sample of Labeled Authentic Data	16
Table 3.2.4: Sample of Labeled Fake Data	16
Table 4.2.1: Before Balancing the Dataset	31
Table 4.2.2: After Balancing the Dataset	31

CHAPTER 1

INTRODUCTION

1.1 Introduction

Fake news, commonly defined as misleading or deceptive content portrayed as news, has the potential to tarnish the reputation of individuals and entities. The proliferation of false information is on the rise, paralleling the exponential growth of social media usage. In today's digital age, social media serves various purposes, including advertising, politics, entertainment, and news dissemination. The unchecked dissemination of intentionally or unintentionally false news poses a significant threat to social media platforms. Individuals, driven by the trend-following behavior and the allure of viral content, often accept information without critical scrutiny. This lack of scrutiny can lead to severe consequences, jeopardizing both individuals and society. A single piece of false information has the power to ruin a person's career, and the rapid spread of false rumors on social media exacerbates this risk. The impact is not limited to online platforms; fake news permeates and contaminates broader societal and cultural contexts. The pervasive influence of social media prompts individuals to take actions based on news from online portals without verifying its authenticity. The consequences are particularly evident in the e-commerce sector, where clickbait tactics deceive consumers into purchasing fake products, damaging the industry's reputation. False information is not restricted to commercial domains; it infiltrates political arenas, where politicians employ fake news to tarnish opponents, potentially inciting conflicts. Innocent citizens become unintended victims in these political battles. While completely eradicating fake news may be challenging, mitigating its spread is within our control. Humans possess the ability to discern falsehoods, leveraging familiarity with news content and common sense to identify satire. A more in-depth investigation can unveil the truth, but the fast-paced nature of modern life often impedes such scrutiny.

To address this issue, my objective is to identify and control the dissemination of false information on online portals and social media channels. Existing efforts in this field involve manual updates by websites that debunk fake news through logical explanations, a time-consuming process. Leveraging deep learning and machine learning methodologies,

our goal is to automate the detection of fake information, reducing its dissemination. Despite over 250 million Bangla language speakers globally, minimal research has focused on detecting Bangla fake news. In Dhaka, where More than 20 lakh Facebook users are currently active., making it one of the world's leading cities in terms of Facebook users, fake news primarily spreads through this type of social media platform.

The objective of this endeavor is to devise a solution for eradicating fake news. This study focuses on constructing an AI-driven, efficient system employing Machine Learning and Deep Learning algorithms to discern and identify fake news. The proposed solution has the following key contributions:

- Here I have created a hybrid model that gives better result compared to other models.
- Conducted a comparative analysis utilizing various feature extraction methods in conjunction with both traditional machine learning and deep learning.
- Established a classification system designed specifically for detecting fake news composed in the Bangla language. Additionally, different pre-trained models have been implemented to address this issue.

1.2 Motivation

In the contemporary era, individuals frequently share information across different social media channels like Facebook, Twitter, and other online sites, as well as traditional news portals. Unfortunately, a prevalent trend involves the dissemination of deceptive Bangla news, leading to these platforms garnering a negative reputation. Articles that initially gain traction are later revealed to be false, resulting in misinformation that misguides readers and gives rise to numerous issues. In our research-focused project, we propose a system where users can conveniently input a news link, and the system will provide an output indicating the probability of the news being fabricated. This approach aims to mitigate confusion surrounding the veracity of news articles. This lack of verification before disseminating news contributed to the escalation of violence. Numerous similar incidents occur daily because of the uncontrolled dissemination of false information.

My research-based project is driven by specific objectives:

1. The swift rise of Bengali fake news on social media underscores the need for an in-depth exploration of future avenues in Bengali fake news detection research to effectively tackle this escalating issue.
2. A thorough review study has revealed distinct patterns that can be leveraged for the detection of false information in Bengali on social media.
3. Individuals often encounter inaccurate information due to misleading headlines, prompting us to conduct this survey with the aim of identifying Bangla fake news.

To address this challenge, I can leverage Machine Learning techniques to develop models capable of automatically detecting fake news. While considerable work has been undertaken in English fake news detection, the progress in Bangla remains limited. Despite the significant number of Bengali speakers globally, comprising approximately 230 million individuals speak Bengali as their first language, and an additional 30 million speak it as a second language, the efforts in detecting Bangla fake news are insufficient. My aim is to make a meaningful contribution to this field and improve the capabilities of false news identification in the Bangla language.

1.3 Rationale of Study

Despite the substantial amount of Bengali language speakers globally, the available resources for research using Natural Language Processing (NLP) are limited. Additionally, there has been minimal effort directed towards fake news detection in this context. The pervasive issue of spreading fake news is particularly pronounced in Bangladesh, leading to frequent chaotic incidents that tarnish the country's reputation. This problem arises as individuals often share misleading information without proper verification.

Addressing this challenge, there is a critical need to create a model with the capability to automatically detect fake news in the Bengali language. While numerous models have been constructed for other languages, there is a noticeable insufficiency when it comes to Bangla language models. A major obstacle in advancing Bangla fake news detection was the lack

of an adequate dataset, which I addressed through meticulous collection and labeling of fake news data for our dataset.

1.4 Research Questions

The objective of this following investigation is to assess the efficacy of a supervised system within the field of news detection. In this context, the supervised system demonstrates commendable performance, while there is room for enhancement in the case of unsupervised systems.

- What types of fake news in consideration?
- What mechanisms are employed in Bangla fake information detection?
- How can Bangla text data be processed within the field of Natural Language Processing (NLP)?
- What will be the methodology for acquiring current and authenticated data?
- What will be the techniques for storing and organizing data?
- What are the Approaches to eliminating routine expiration from the dataset?
- What is the overview of recent endeavors in the field of Bangla false news detection?
- What will be the future work?

1.5 Expected Output

The entire research revolves around constructing a model designed to enhance the accuracy of news articles. In this system, the input consists of news title, category, description, and link. Subsequently, built-in functions are employed to eliminate various varieties of disturbances are present, and the news is subjected to evaluation through our model. This model generates results by comparing the keywords of fake and real news, storing relevant keywords for each category. The storage of keywords is achieved through the training of datasets containing both fake and real news. Hence, through identifying keywords that signal either fake or genuine news, our model delivers a conclusive determination or outcome regarding the news's credibility.

This research endeavor is focused on achieving the following objectives:

1. Classifying Bangla news into a binary format to determine its authenticity.
2. Comparing true and false news to ascertain the likelihood of a news article being fake.
3. Enhancing the accuracy of Bangla fake news detection.

However, we can develop a model utilizing NLP techniques to automatically identify fake news, thereby mitigating its impact.

1.6 Finance and Project Management

There isn't a financial aspect involved since all the data collection and associated tasks are personally undertaken by me.

1.7 Report Layout

The report structure provides an overview of all the chapters. The summaries of each chapter are as follows:

Chapter 1: Covers Introductory part as well as Motivation behind the work and also about the project management and also about the output expectations.

Chapter 2: Explores information's about the work as well as the other related work and the problem statements as well as the challenges.

Chapter 3: Details about the Instruments and mostly The Data Collection Process Also Analysis, The Methodology I proposed and The Requirements for implementations.

Chapter 4: Presents Setup for the experiment, Results, Evaluation and Discussion.

Chapter 5: Examines Environmental Impact, Social Impact, Ethical Considerations, Sustainability Strategy.

Chapter 6: Provides a Study Recap, Concluding Remarks, and Implications for Future Research.

CHAPTER 2

BACKGROUND

2.1 Terminologies

In the current Digital communication era., the dissemination of fake news has become more sophisticated than ever before. Nowadays, numerous mainstream online news outlets engage in deceptive practices for various reasons. Unfortunately, our diminishing moral values and ethical standards contribute to a lack of concern for the potential consequences that fake news may have on our society. Consequently, when one fake news site is banned, another emerges to take its place, perpetuating the cycle. Additionally, news originating from a specific source may lose its clarity through continuous sharing and resharing by individuals. Furthermore, verifying the authenticity of news becomes challenging, especially when shared by close acquaintances.

Moreover, in the contemporary world, everything is centered around the digital platform, heavily dependent on the internet. People no longer make the effort to purchase newspapers or invest substantial time clicking on links to read full news articles. Instead, they simply scroll through social media and glance at the headlines. This casual approach often results in misinformation and misjudgment. In our research-based project, we conducted a survey to identify misleading headlines using a machine learning model. We compiled datasets for both true and false cases. Through a comparative analysis of these two scenarios, our model utilizes a decision tree to provide an accurate determination of whether the news is true or fake.

2.2 Related Works

The dissemination of fake news poses a noteworthy threat to our well-being. Despite the advantages of modern technology, it has the potential to mislead us through various channels. The increasing engagement of individuals in online activities makes it easier to reach a broader audience within seconds. Presently, people frequently share content on social media without proper authentication, impacting both the online and physical realms. In some instances, individuals take actions based on information they encounter online. To curb the dissemination of false information, it is imperative to identify and prevent it

automatically. Numerous studies have been conducted on the identification of false news in the English language, yet there is a noticeable dearth of research focused on Bangla fake news detection. Undeterred by this gap, our initiative involves leveraging machine learning models' and (NLP) natural language processing to address this concern. In our pursuit to mitigate the proliferation of fake news, we conducted a comprehensive survey. Collecting data from social media encompassing both true and false narratives, we meticulously curated and trained our dataset. By employing a machine learning model, we engaged in a comparative analysis of these divergent news types, enabling the identification of falsehoods and truths. To facilitate this discernment, we constructed a binary classifier utilizing deep learning methodologies.

Our research draws inspiration from previous studies. In [1], fake news detection involved utilizing linguistic features and methods based on neural networks., achieving a best result of 91 F1-score with SVM on an 8.5k manually labeled dataset. In [2], MNB and SVC classifiers were used for Bangla fake news detection, achieving 96% accuracy with Linear kernel SVM on a 2.5k dataset. [3] proposed a model using various classifiers on a dataset comprised of 3.5k authentic data and 2.3k fake data, with the Passive Aggressive Classifier and Support Vector Machine achieving accuracy rates of 93.8% and 93.5%, respectively. In [4], a machine learning-based model achieved 84.57% accuracy on a dataset of 7000 Bengali text documents using The SGD classifier, utilizing 'tf-idf' with a combination of unigram and bigram features, was employed in the analysis. [5] utilized LSTM and CNN methods, achieving 78% accuracy on a dataset with approximately 50k news. [6] A model was constructed using Bag of Words, TF-IDF matrix, and a Random Forest Classifier on a dataset of 500 news, achieving 86% accuracy. [7] Employed hybrid deep learning models that combine Convolutional Neural Networks (CNN) with Machine Learning classifiers, achieving 99% accuracy. In [8], a Machine Learning-based model on a 6.5k news dataset achieved the best accuracy of 94%. The rise of distorted news and "alternate facts" has become a societal concern, notably accelerated by the term "fake news," popularized by US President Donald Trump. The spread of such news during the 2016 US presidential election raised criticism against platforms like Facebook. Reality-checking approaches,

relying on automated verification and structured queries to databases, have been used to assess the credibility of news claims. However, challenges exist, such as the availability of external sources and the effectiveness of these approaches in detecting misinformation in texts lacking visible information. Additionally, efforts have been made in the automatic identification of deceptive content, exploring various domains like forums, consumer review sites, online advertising, online dating, and crowdfunding platforms. While fake news detection shares similarities with deception detection, they differ in fundamental aspects.

Comparative Analysis: Here we have a comparison with other related studies. Here the comparative analysis tables show the other related work with the dataset, used models and their best result.

TABLE 2.2.1: Comparative analysis

Work	Dataset	Used Models	Best Accuracy
This Paper	Authentic Data 48678, Fake Data 1299	Bidirectional GRU Model, Convolutional Neural Network (CNN) + Long Short-Term Memory (LSTM) + GlobalMaxPooling1D layers (Hybrid Model), Long Short-Term Memory (LSTM), and 1D Convolutional Neural Network (CNN).	99.13%
Hussain, Md Gulzar, et al. (2020)	Real Data 1548, Fake 993	Multinomial naive bayes and Support vector machine	96.64%
Sharif, Omar, et al (2020)	7115 texts among them 3557 texts	LR, DT, RF, MNB, SGD	84.57%

	are suspicious, and 3558 texts are non-suspicious.		
Mahabub, A. (2020)	Data Set of 6500 data from which about 3252 data are fake, and 3259 data are real	K-Neighbors, Ada Boost, Decision Tree, Random Forest, Extra Tree, SVC, Gradient Boosting, Logistic Regression, Multi-Layer Perception (MLP), Multinomial Naïve Bayes, X-Gradient Boosting	94.5%
Akter, Farhana, et al. (2021)	The dataset contains 2046 data.	Multinomial naive Bayes, K-nearest neighbor, Random Forest Classifier, Decision tree classifier, Support vector machine, Logistic Regression	95%
Rasel, Risul Islam, et al. (2022)	The dataset contains 4678 news data.	Machine Learning (LR, SVM, KNN, MNB, Adaboost, and DT), Deep Neural Networks (LSTM, BiLSTM, CNN, LSTM-CNN, BiLSTM-CNN)	95.9%
Islam, Farzana, et al. (2020)	Dataset of 726 articles	Naive Bayes, Logistic Regression, Random Forest	85%
Mugdha, et al. (2020)	The dataset contains 538 News Data	Support Vector Machine (SVM), Logistic Regression (LR), multilayer perceptron (MLP), Random Forest Classifier (RF),	87.42%

		Voting Ensemble Classifier (VEC), Gaussian Naive Bayes (GNB), Multinomial Naive Bayes (MNB), AdaBoost (AB) and Gradient Boosting (GB).	
--	--	--	--

2.3 Summary

In the investigation, we employed both machine learning as well as deep learning methodologies for the detection of Bangla fake information's. A diverse range of deep learning models was utilized in our study. To facilitate model training, we meticulously curated an extensive dataset through manual collection, encompassing various news and social media platforms. During the initial phase of data collection process, we extracted Bangla news information's from diverse media sources and subsequently synthesized individual information into eight distinct sections, including date, source, domain, relation, headline, article ID, label, and content. The total dataset comprised over fifty thousand entries. Prior to implementing any deep learning methods, we subjected the entire Bangla dataset to preprocessing. In this preprocessing phase, I diligently identified and removed noise. Following the elimination of unwanted elements, I applied a count vectorizer to transform the text into vector-based frequencies for each word within the content. Subsequently, I conducted multiple training iterations for our model, culminating in highly favorable outcomes.

2.4 Scope of Problems

While the identification of fake information's in Natural Language Processing is commonplace, the development of Bengali NLP remains in its early stages. My research focuses on the application of machine learning and deep learning techniques for identification purposes. The presence of a substantial amount of noise within our dataset posed a significant challenge, as it hindered the attainment of the desired results. Simultaneously, a critical issue arose concerning the scarcity of fake news instances in our

data. Efforts to address this problem included augmenting the dataset with additional fake data, although this proved insufficient. Consequently, I took measures to incorporate a proportion of authentic data. During this research, a combination of deep learning and machine learning methods was employed, yielding highly satisfactory results.

2.5 Challenges

Initiating work on this topic presented the most significant challenges I have encountered thus far. The primary setback lies not in the procedure itself but in the scarcity of resources. Due to the limited work done with the Bangla language, there is a paucity of datasets available. Compounding the challenge was the difficulty in data collection, as fake and real news lacked distinct and easily identifiable characteristics. Consequently, I had to collect data from various sources like online resources, newspaper, Kaggle and other resources, introducing challenges related to maintaining data validity and authenticity. Furthermore, the acquired data lacked proper structure and organization, adding complexity to the process. The labeling phase demanded heightened attention to detail. In the preprocessing stage, we implemented fresh coding steps to align with the model, addressing challenges in running the dataset, which included multiple breakdowns in the device. The Bangla dataset, presented as a string, posed additional challenges. In contrast to the well-organized datasets available for the English language, Bangla language resources are still in their infancy. The lack of supervised materials necessitated our independent efforts. Collecting fake news emerged as another major issue, with our aim being to gather as much fake data as possible to enhance the precision of the results.

CHAPTER 3

Research Methodology

3.1 Instrumentation and Research Subject

To achieve our objectives, we implemented various procedures, which we will elaborate on in this chapter. Striving for optimal results, we experimented with diverse approaches involving both machine learning as well as deep learning algorithms. Our model underwent training using various classifiers, including K-Nearest Neighbor, Multinomial Naive Bayes, Support Vector Machines, and sophisticated models like Bidirectional GRU Model, Convolutional Neural Network (CNN) + Long Short-Term Memory (LSTM) + GlobalMaxPooling1D layers (Hybrid Model), Long Short-Term Memory (LSTM), and 1D Convolutional Neural Network (CNN). Prior to training, data preprocessing was imperative due to the prevalent noise in the dataset. Given the nature of our work in NLP (Natural Language Processing), the significance of a robust dataset cannot be overstated. While we initially collected a dataset online, its adequacy was limited, prompting us to manually gather additional data to construct a substantial dataset. This chapter provides a detailed account of the various classifiers employed in our research.

The undertaking was extensive, utilizing a blend of machine learning and deep learning classification techniques. Executing our programs for optimal results necessitated rapid and frequent runs, a process that proved time-consuming, particularly with limited hardware resources. To expedite our workflow, we leveraged the following software and hardware components like: 256GB SSD, Intel Core i5 processor, Google Colab, 8GB RAM.

3.2 Data Collection Procedure

Data serves as the cornerstone of any research endeavor, and its collection stands as the foundational and paramount step in the research process. The primary goal of data collection is to ensure the reliability of the information being gathered. A robust dataset is instrumental in achieving more accurate research outcomes. Given that the dataset constitutes a crucial aspect of our research, we must ensure the availability of a substantial dataset before commencing our work.

Our focus revolves around Bangla fake news detection using NLP (Natural Language Processing). Due to the scarcity of Bangla datasets, we resorted to manual data collection. While we obtained a dataset online, its adequacy was insufficient. Our dataset comprises both Bangla authentic news and Bangla fake news, totaling more than 50k entries. I have started collecting data with very proper planning. Firstly, I identified the data sources. Then finalized the data type. We are living in an age of data. Lots of have been created every moment. So, there is a vast amount of news created every day all over the world. As I am working with Bangla data, I have selected various sources of Bangla data. Mostly the data is from Kaggle (BanFakeNews: A Dataset for Detecting Fake News in Bangla). The dataset is segmented as 48k Authentic Data, 7k labeled authentic data, 1k fake data, and 1k fake labeled data.

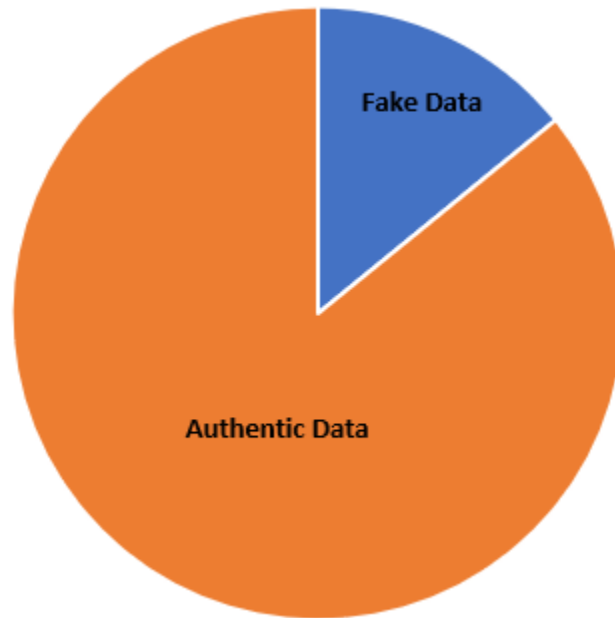


Figure 3.2.1: Ratio of authentic data and fake data

Here in the graph, we can see the ratio of Authentic and Fake data. We have collected both fake and authentic data to train our model with both type of data so that the model gives more accurate result.

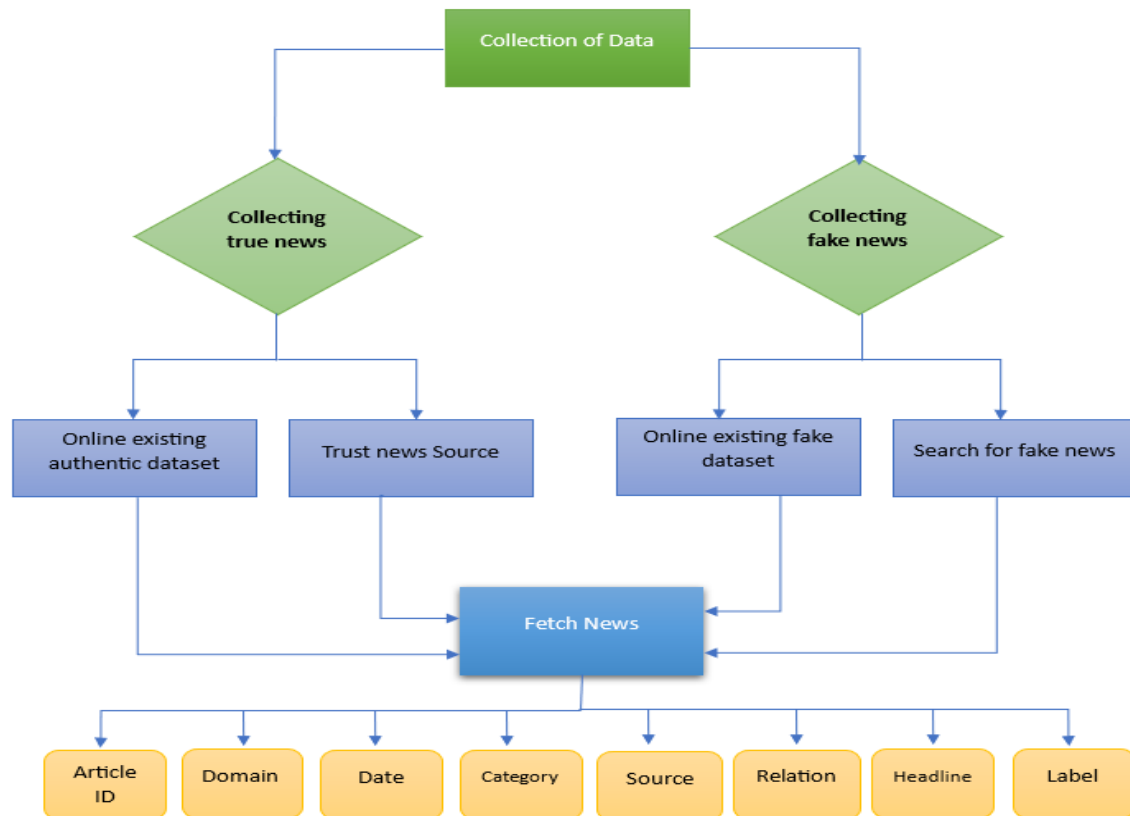


Figure 3.2.2: Data Collection Workflow

Upon establishing a well-structured dataset, our work commenced. The initial phase involved the preprocessing of our dataset, a crucial step in data manipulation aimed at enhancing model performance. Given that a substantial portion of the dataset contained noisy data, preprocessing was deemed essential to facilitate the smooth creation of a model. The steps undertaken for dataset preprocessing are outlined below.

Punctuation removal: Similar to the English language, the Bengali language incorporates numerous punctuation characters in its text. However, these punctuation characters are irrelevant for our dataset. Consequently, we eliminated these characters from the data using appropriate libraries.

Tokenization: Data tokenization involves breaking down a sentence, phrase, paragraph, or an entire text document into individual words or tokens. Tokenization is a crucial preprocessing step that involves filtering out the essential words.

Stopword removal: In every natural language, stop words are the most common words. When analyzing text data and constructing NLP models, these stop words usually add minimal meaning to the document. Stop words typically include articles, conjunctions, prepositions, and specific pronouns, such as 'on,' 'a,' 'the,' 'an,' 'the,' 'but' etc. Similarly, the Bengali language contains its set of stop words that must be removed from the dataset to enhance processing efficiency.

Sample Authentic Data Table: Here in the dataset, I have 48678 rows and 7 columns like articleID, date, domain, headline, category, content, and label. I have labeled the authentic data as a value of 1. Here 1 indicates authentic news data.

Table 3.2.1 Sample of Authentic Data

articleID	domain	date	category	headline	content	label	
0	1	jagonews24.com	2018-09-19 17:48:18	Education	হট্টগোল করায় বাকুবিতে দুইজন বরখাস্ত, ৬ জনকে শোকজ	গত ১৭ সেপ্টেম্বর বাংলাদেশ কৃষি বিশ্ববিদ্যালয়ে ...	1
1	2	jagonews24.com	2018-09-19 17:48:19	National	মালয়েশিয়ায় কর্মী পাঠানোর ব্যবস্থা নেয়ার সুপারিশ	বাংলাদেশের বৃহৎ শ্রমবাজার মালয়েশিয়ায় আবার শ্রম...	1
2	3	jagonews24.com	2018-09-19 17:48:20	National	প্রেমের প্রস্তাবে রাজি না হওয়ায় কুলছাত্রীকে ...	নরসিংদীর মনোহরদীতে প্রেমের প্রস্তাবে রাজি না হ...	1
3	4	jagonews24.com	2018-09-19 17:48:21	Crime	মেডিয়েশনই মামলাজট নিরসনের পথ : বিচারপতি আহমেদ ...	সুপ্রিম কোর্টের হাইকোর্ট বিভাগের বিচারপতি আহমেদ...	1
4	5	jagonews24.com	2018-09-19 17:48:21	National	টকশোতে বক্তব্য দিতে গিয়ে জাপা নেতার মৃত্যু	মাদারীপুর সদরের উপজেলার লোকেরপাড়ে একটি বেসরকার...	1

Sample Fake Data Table: Then there is a fake data segment of 1299 rows and 7 columns like articleID, date, domain, headline, category, content, and label. I have labeled the fake data as a value of 0. Here 0 indicates fake news data.

Table 3.2.2 Sample of Fake 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	সেপ্টেম্বর ১৭, ২০১৮	Miscellaneous	বিগিভিতে যেবার আমি ইন্টারভিউ দিতে গেলাম	BTV থেকে লোকজন আসছে, ইন্টারভিউ নিবে।চারজনের টি...	0
2	3	earki.com	২০:০৯, জানুয়ারি ১৪, ২০১৯	Miscellaneous	বিদেশ থেকে উন্নতমানের বি-রোবীদল আমদানি করার পরা...	অদ্ভুত বিরোধীদলহীনতায় ভুগছে সরকার। এ এক অন্যরক...	0
3	4	channelhdhaka.news	2018-06-30T15:56:47+00:00	Sports	অবসর নেয়ার ঘোষণা দিলেন মেসি।	রাশিয়া বিশ্বকাপ নকআউট পরে ফ্লোরের সাথে ৪-৩...	0
4	5	motikontho.wordpress.com	2013-03-05T21:55:45+00:00	Miscellaneous	মাদারফাকার নহে, ব্রাদারফাকার: সাকা। দৈনিক মতি...	নিজস্ব মতিবেদক মাদারফাকার নহে, আমি ব্রাদারফাকা...	0

Sample Labeled Authentic Data Table: Then there is labeled authentic news data of 7202 rows and 9 columns like articleID, domain, source, date, headline, category, content, relation, and label.

Table 3.2.3 Sample of Labeled Authentic Data

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 বিক্রেতা বিশেষ অভিযান পরিচালনা কর...	1.0
3	4	bd24live.com	2018-09-19 18:27:56	National	জননিরাপত্তা বিভাগের সচিব	Related	উক্রানি রোহে নজরদারি থাকবে সামাজিক যোগাযোগ মাধ...	সনাতন ধর্মাবলম্বীদের সবচেয়ে বড় ধর্মীয় উৎসব দুর্...	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

Sample Labeled Fake Data Table: There is also labeled fake data of 1202 rows and 9 columns like articleID, domain, source, date, headline, category, content, relation, label and F-type.

Table 3.2.4 Sample of Labeled Fake Data

articleID	domain	date	category	source	relation	headline	content	label	F-type	
0	1	channelhdhaka.news	2019-03-14T13:34:14+00:00	International	Reporter	Unrelated	মুরগির হামলায় শেয়াল নিহত	বাংলায় একটা প্রবাদ আছে, শেয়ালের কাছে মুরগী বর্...	0	Satire
1	2	earki.com	সেপ্টেম্বর ১৭, ২০১৮	Miscellaneous	Reporter	Unrelated	বিটিভিতে খেবার আমি ইন্টারভিউ দিতে গেলাম	ETV থেকে লোকজন আসছে, ইন্টারভিউ নিবে। চারজনের টি...	0	Satire
2	3	earki.com	২০১৯, জানুয়ারি ১৪, ২০১৯	Miscellaneous	Reporter	Unrelated	বিদেশ থেকে উন্নতমানের বিরোধীদল আমদানি করার পরা...	অদ্ভুত বিরোধীদলীয়নতায় ভুগছে সরকার। এ এক অন্যরক...	0	Satire
3	4	channelhdhaka.news	2018-06-30T15:56:47+00:00	Sports	Reporter	Unrelated	অবসর নেয়ার ঘোষণা দিলেন মেন্সি।	রাশিয়া বিশ্বকাপ নকআউট পরে ফ্রান্সের সাথে ৪-৩...	0	Satire
4	5	motikonho.wordpress.com	2013-03-05T21:55:45+00:00	Miscellaneous	Reporter	Unrelated	মাদারফকার নখে, ব্রাদারফকার: সাকা দৈনিক মতি...	নিজস্ব মতিবদক মাদারফকার নখে, আমি ব্রাদারফকা...	0	Satire

All over the dataset most of the data is collected from various sources of news. Here in the dataset, there is a mix of various category data like international, national, sports, entertainment, crime, etc. I have collected and chosen this dataset very carefully so that I can analyze the dataset very well and perform the study properly.

3.3 Statistical Analysis

We initiated the preparation process by eliminating duplicate information from our datasets. The subsequent step entailed the exclusion of emotions, numerals, links, user tags, capitalization, extended words, Uniform Resource Locator (URL), and user mentions from all three datasets. To ensure precision in our findings, sentences that combined Bangla and Romanized Bangla were specifically excluded from the dataset. Additionally, since the

messages were primarily in the native language, we opted not to use stop part-of-speech tagging or stemming on our data.

In our dataset, there are more than 50k organized text data collected from various online media and websites. The dataset was formatted in Excel, with the file extension being .xlsx. Here we have 49977 train data.

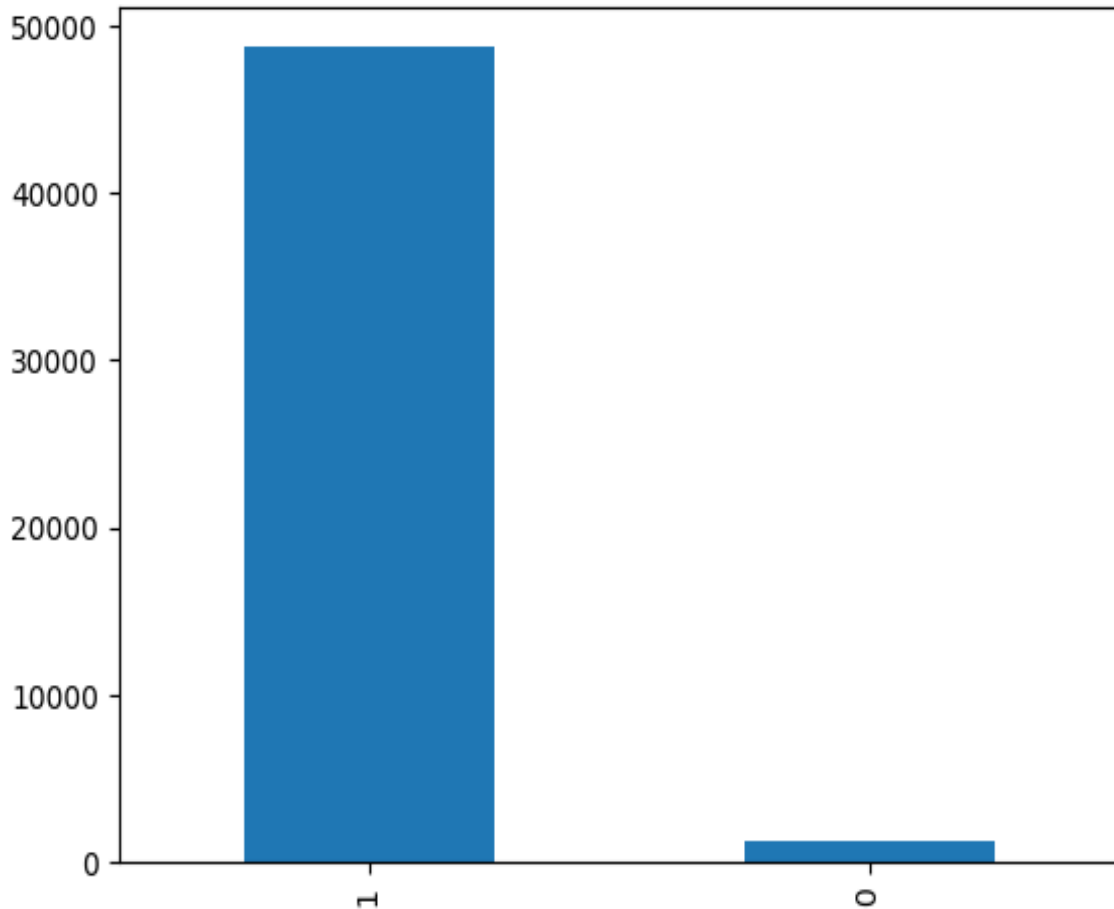


Fig 3.3.1 Graphical representation Train data

This visual representation utilizes binary numbers (1 for authentic and 0 for fake) to illustrate the proportion of authentic and fake data in a training dataset. It offers a graphical overview of class distribution, facilitating the evaluation of dataset balance and identification of potential biases. Examining this ratio is essential for optimizing machine

learning model training, particularly in mitigating imbalances between classes to enhance generalization and performance. Also, there is 8501 test data.

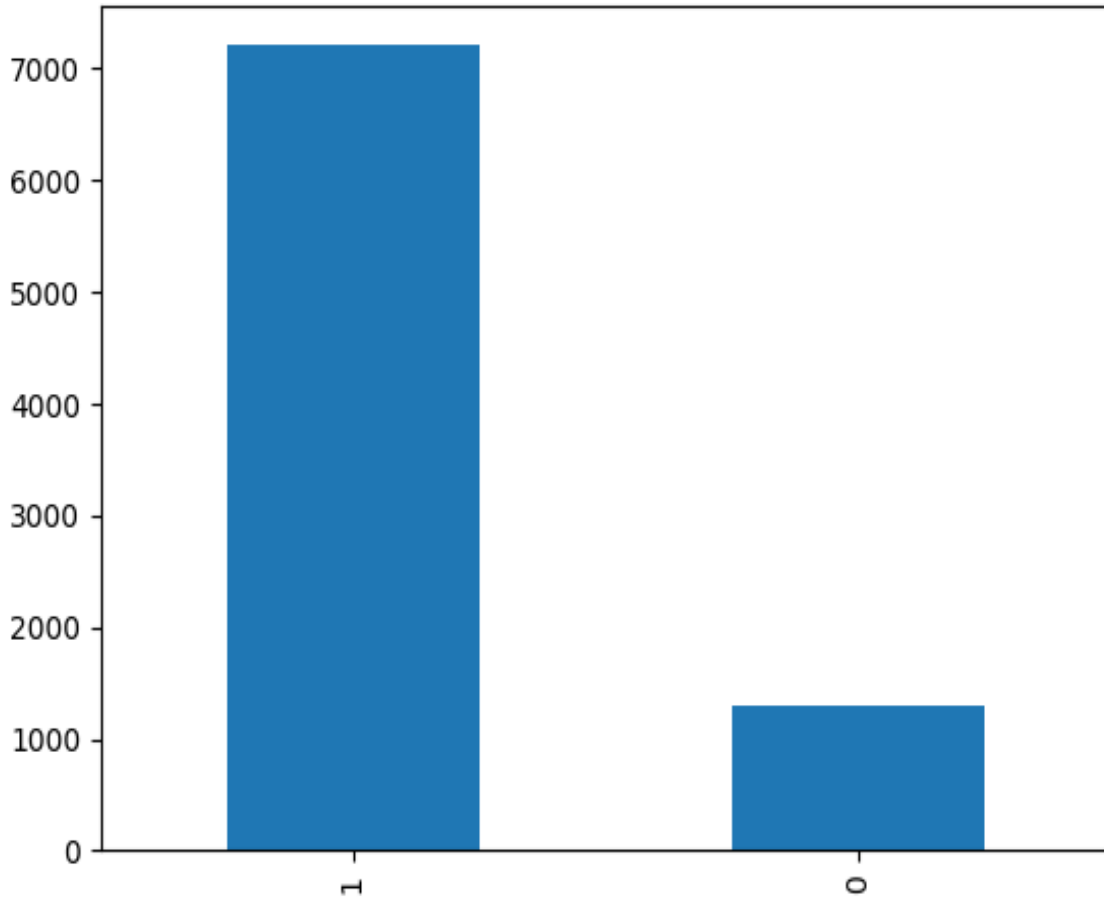


Fig 3.3.2 Graphical representation Test data

In the context of 8,501 test data points, a visual representation is utilized to depict the ratio of authentic and fake data. The binary encoding system is applied, where 1 denotes authentic data and 0 signifies fake data. This graphical representation offers a succinct overview of class distribution in the test dataset, facilitating the evaluation of balance and potential biases. Examining this ratio is crucial for assessing the generalization capability of a machine learning model trained on a specific dataset, particularly in distinguishing between authentic and fake instances when exposed to unseen data.

NLP Images: Here, we present the output of the dataset with keyword visualization. To generate this image, we make use of the word cloud analysis library function. Initially, we gather keywords by employing the word tokenize function, followed by arranging the keywords based on their frequency using the sorted function. Drawing the image involves utilizing functions such as plt.figure, plt.axis, plt.imshow, and plt.tight_layout. Finally, the image is displayed using the plt.show() function.



Figure 3.3.3: Train Plot NLP Image

Train Plot NLP image: Figure 3.7 displays certain keywords extracted from our model corresponding to Train Plot. In this representation, the more highlighted keywords indicate those that the machine identified with the highest frequency.



Figure 3.3.4: Test Plot NLP Image

Test Plot NLP image: Figure 3.8 illustrates specific keywords obtained from our model for Test Plot. In this depiction, the more highlighted keywords signify those for which the machine recorded the highest frequency.

3.4 Proposed Methodology/Applied Mechanism

Bidirectional GRU Model

The Bidirectional GRU (Gated Recurrent Unit) stands as a variant of the recurrent neural network (RNN) specifically crafted for sequence modeling. This architecture engages with input sequences in both forward and backward directions, adeptly gathering context from both preceding and subsequent states. Critical elements encompass GRU layers featuring gating mechanisms, embedding layers for token representation, dropout to ensure

regularization, and dense layers for generating predictions. This structural design proves highly effective in tasks demanding a thorough grasp of contextual nuances, such as sentiment analysis or named entity recognition. The bidirectional nature significantly augments the model's capability to comprehend intricate sequential patterns.

This is my proposed model. This gives the best result in this study. Here the Confusion Matrix and ROC Curve shown below for this study with Bidirectional GRU Model.

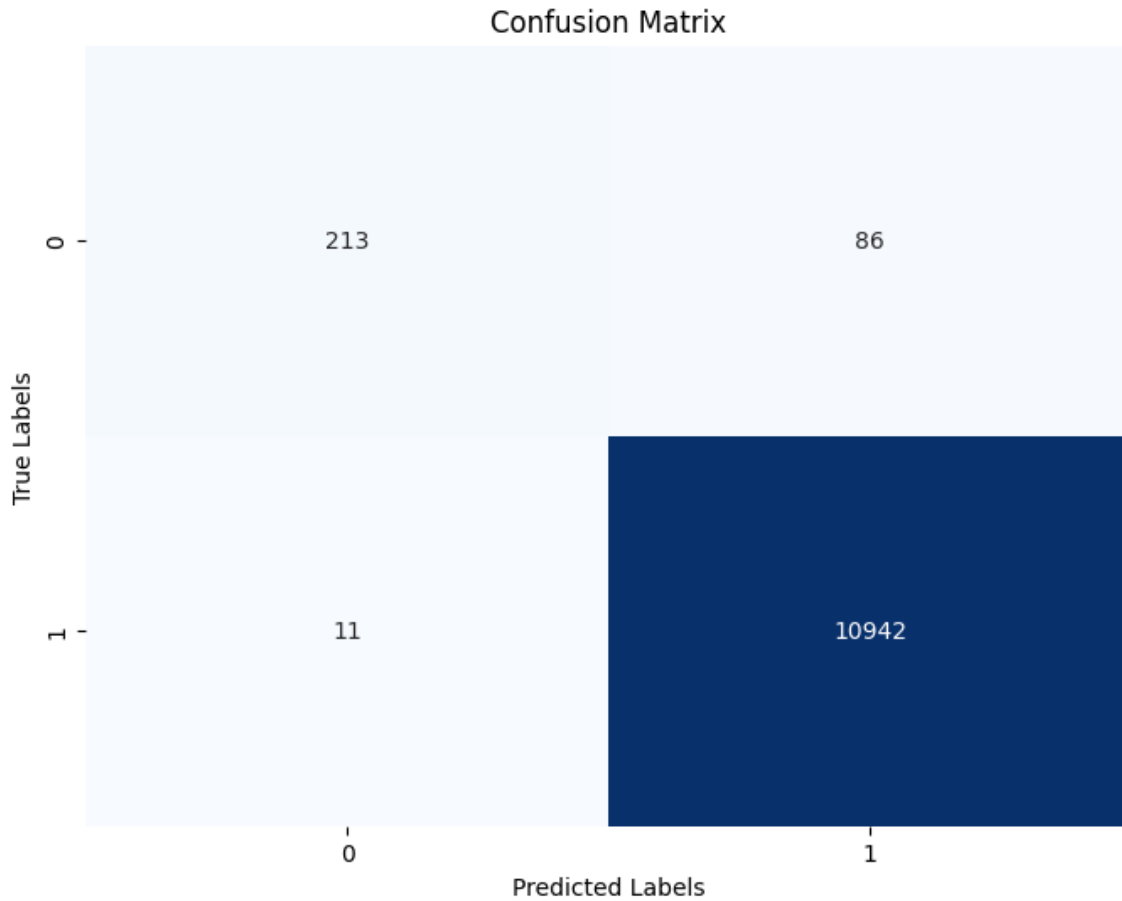


Fig 3.4.1 Confusion matrix

The Bidirectional GRU model demonstrates outstanding performance, achieving an accuracy of 99.13%, precision of 99.22%, recall of 99.89%, and an F1 score of 99.55%. These metrics emphasize the model's robust capability to precisely classify instances of both authentic and fake data.

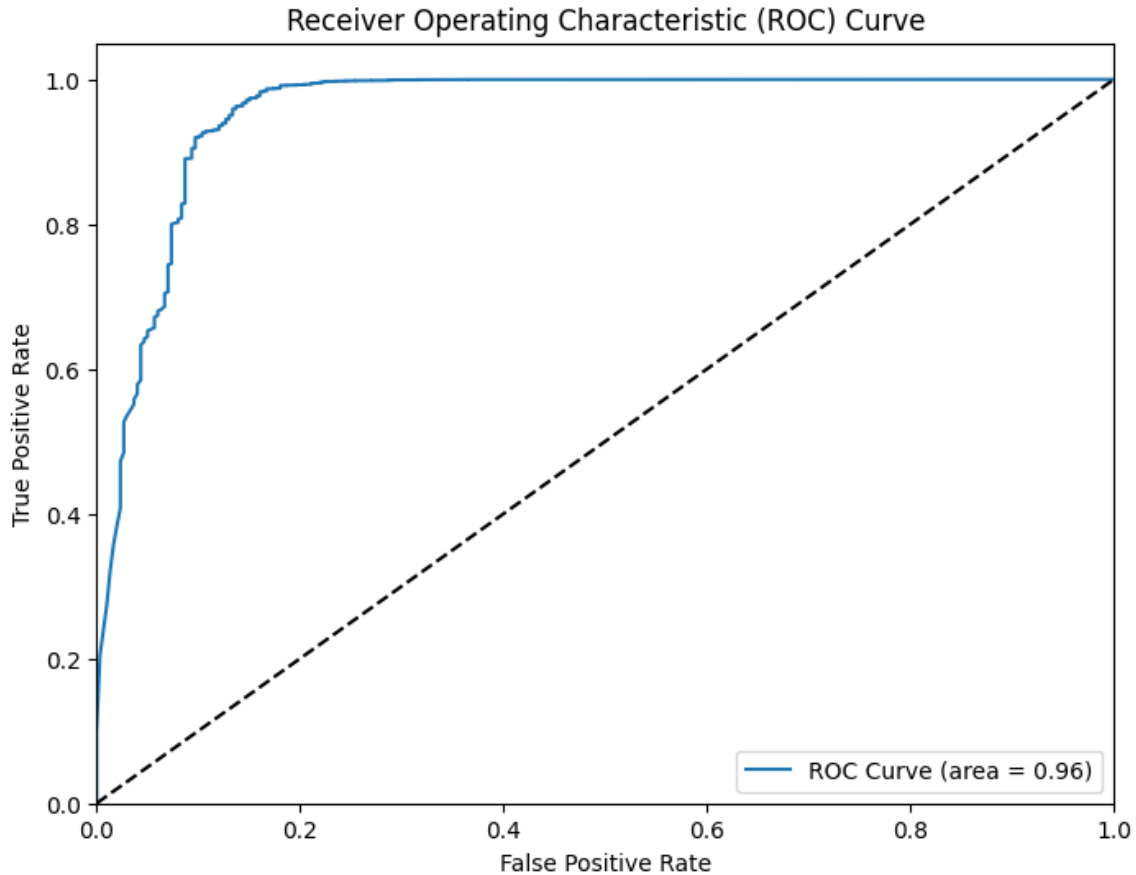


Fig 3.4.2 ROC Curve

1D Convolutional Neural Network (CNN)

The 1D Convolutional Neural Network (CNN) is an architectural framework within neural networks tailored for processing sequential data, such as time series or text sequences. Diverging from traditional CNNs designed for two-dimensional data like images, 1D CNNs operate specifically on one-dimensional sequences. Using convolutional layers, they autonomously acquire hierarchical representations of patterns and features inherent in the input data. These networks exhibit notable effectiveness in tasks like text classification, speech recognition, and signal processing, where the ability to capture local dependencies and patterns in sequential data holds paramount importance. The 1D CNN architecture comprises convolutional layers, and pooling layers for down-sampling, and frequently integrates fully connected layers to derive predictions based on the learned features.

Here the Confusion Matrix and ROC Curve shown below for this study with 1D Convolutional Neural Network (CNN)

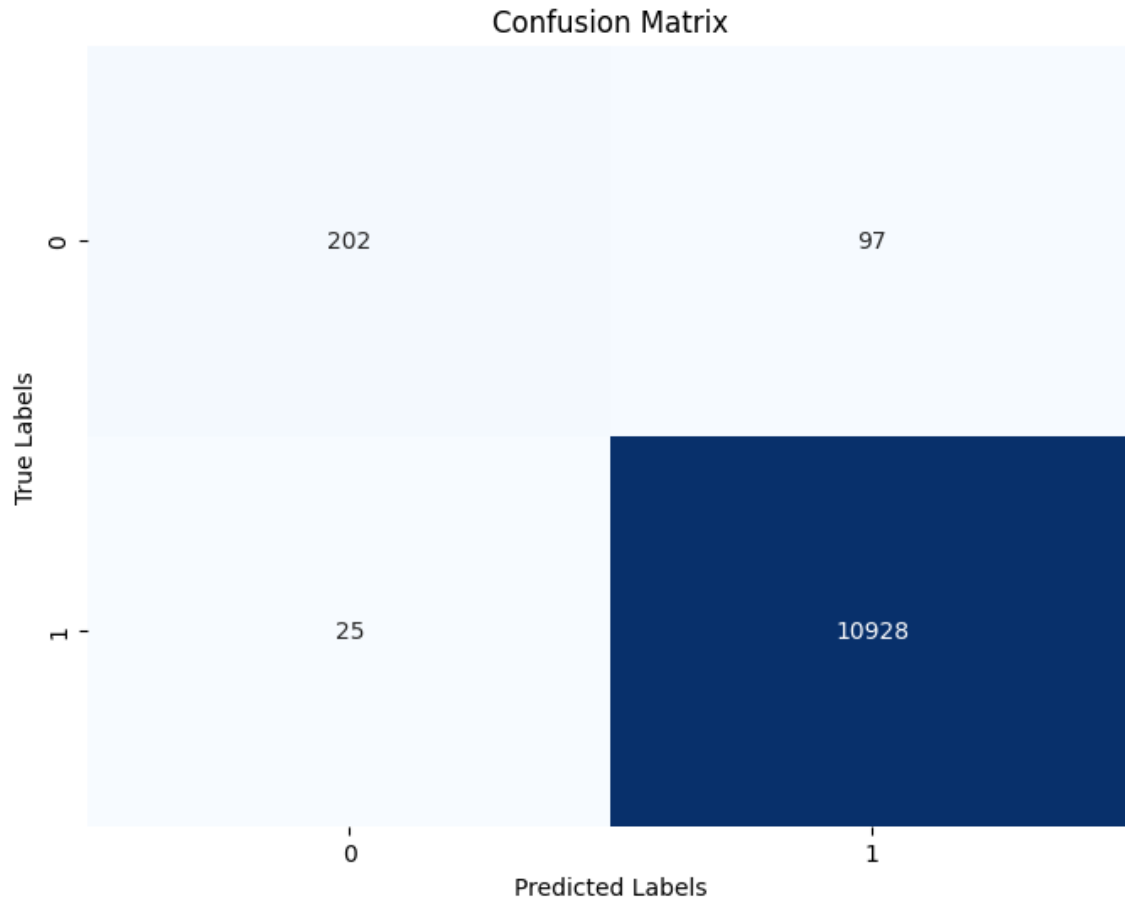


Fig 3.4.3 Confusion matrix

The 1D Convolutional Neural Network (CNN) has exhibited remarkable performance, attaining an accuracy of 98.91%, precision of 99.12%, recall of 99.77%, and an F1 score of 99.44%. These metrics underscore the model's strength in precisely categorizing data, demonstrating high precision, recall, and an overall robust ability to distinguish between authentic and fake instances.

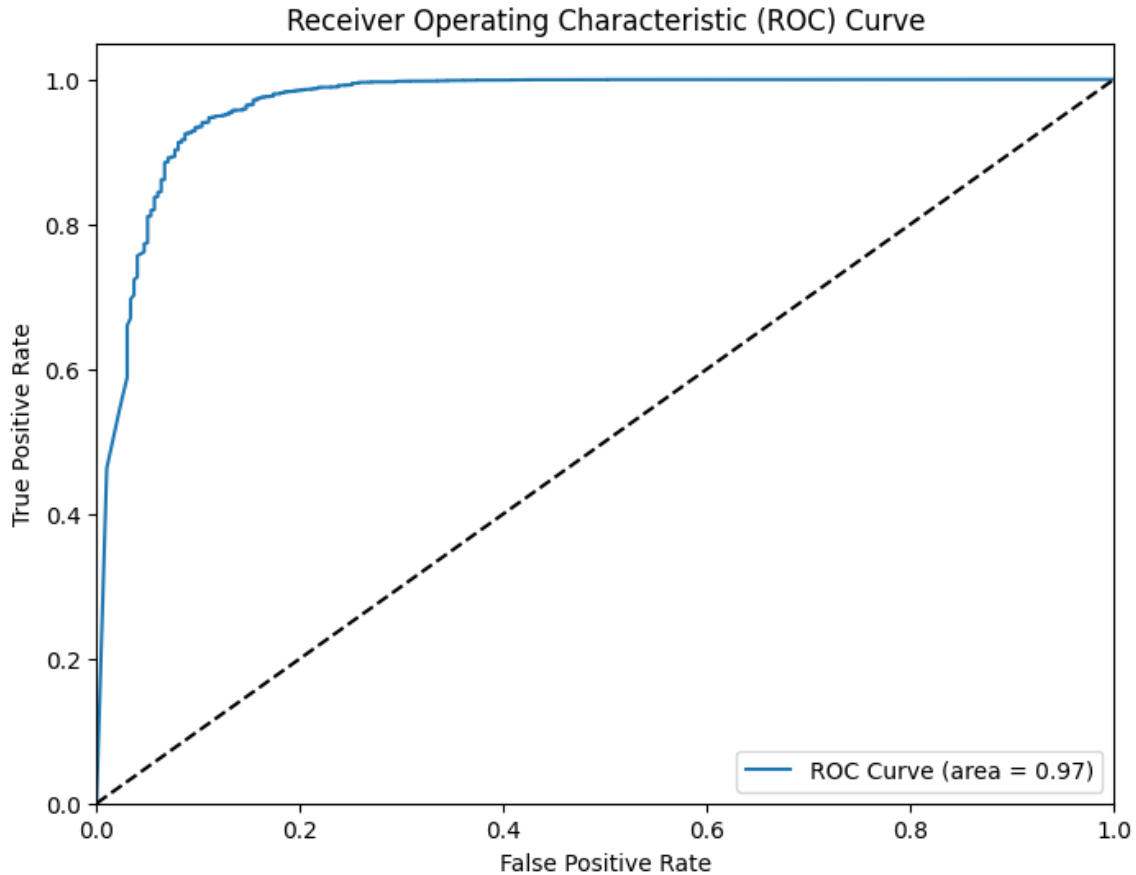


Fig 3.4.4 ROC Curve

Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) constitutes a specific architecture within recurrent neural networks (RNNs), meticulously crafted to address the vanishing gradient problem inherent in conventional RNNs. This design breakthrough empowers LSTMs to adeptly process long-range dependencies in sequential data. The core of LSTMs lies in a memory cell equipped with self-regulating gates—namely, input, forget, and output gates—that meticulously manage the information flow within the cell. This distinctive architecture grants LSTMs the capability to selectively retain or discard information over extended sequences, rendering them remarkably effective in tasks spanning time-series data, natural language processing, and speech recognition. The unique proficiency of LSTM to capture and retain information over prolonged periods sets them apart, rendering them particularly

well-suited for applications demanding an understanding of context and dependencies across distant elements in a sequence.

Here the Confusion Matrix and ROC Curve shown below for this study with LSTM (Long Short-Term Memory)

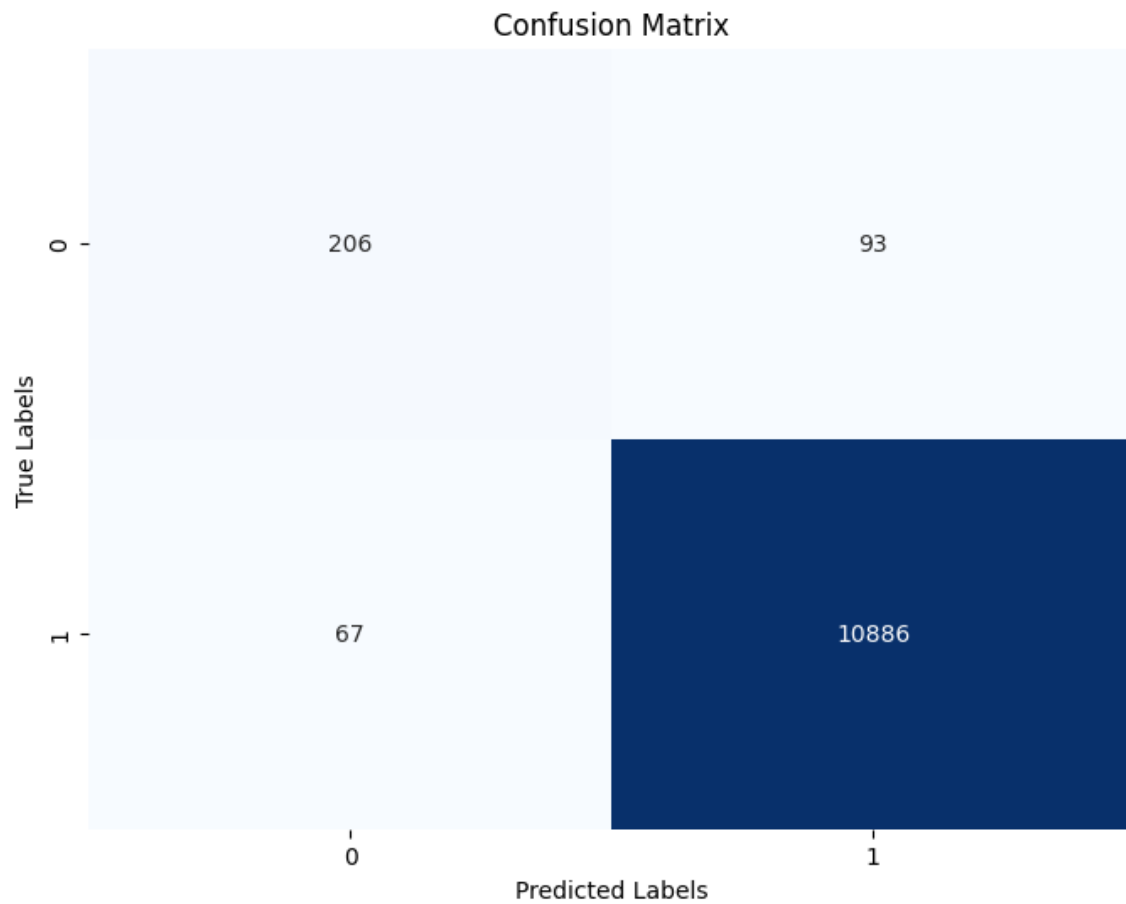


Fig 3.4.5 Confusion matrix

The Long Short-Term Memory (LSTM) model has exhibited commendable performance, achieving an accuracy of 98.57%, precision of 99.15%, recall of 99.38%, and an F1 score of 98.27%. These metrics highlight the model's robust capability to accurately categorize data, demonstrating high precision, recall, and an effective overall balance between false positives and false negatives in distinguishing between authentic and fake instances.

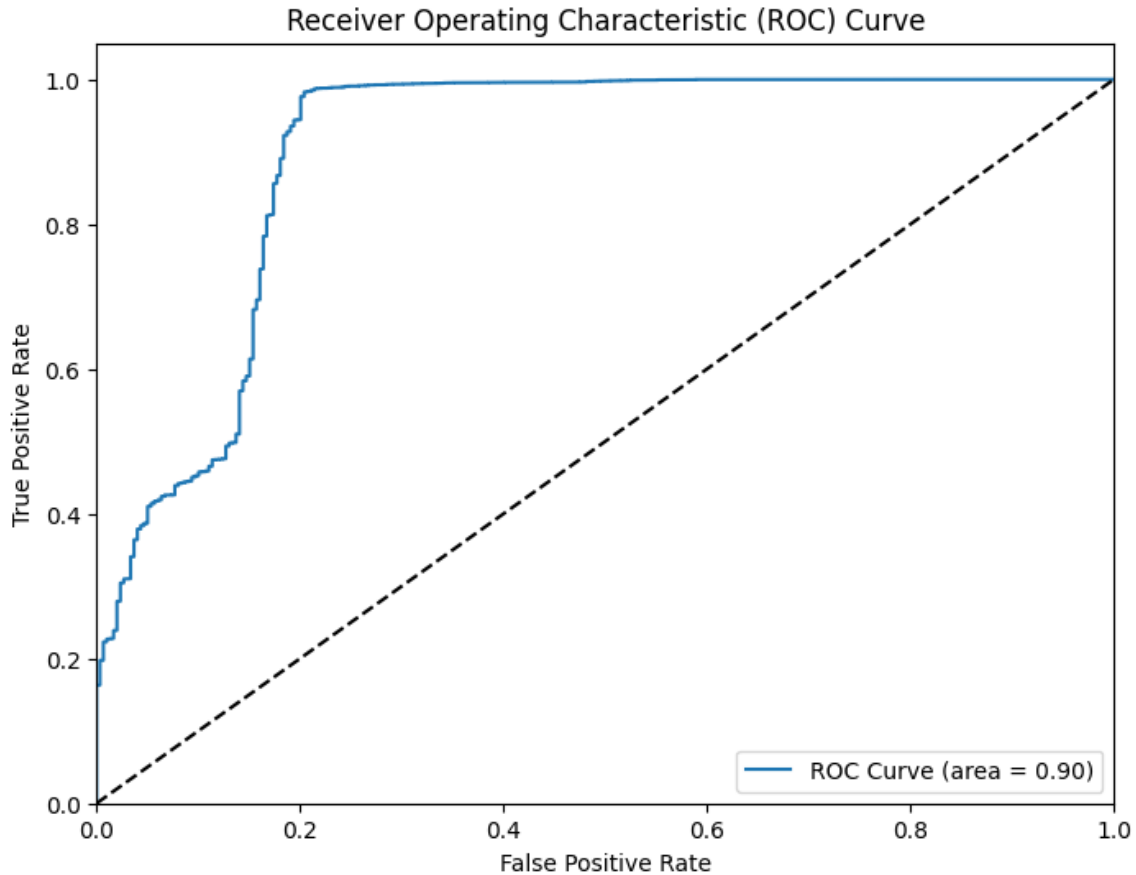


Fig 3.4.6 ROC Curve

Convolutional Neural Network (CNN) + Long Short-Term Memory (LSTM) + GlobalMaxPooling1D layers

This is my self-created model. The fusion model incorporates three fundamental layers—Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and GlobalMaxPooling1D. The CNN layer excels in capturing spatial patterns, while the LSTM layer tackles sequential dependencies, proving especially beneficial for time-series data or sequences. The GlobalMaxPooling1D layer complements the architecture by extracting the most significant features. This amalgamation capitalizes on the strengths of both CNN and LSTM, amplifying the model's proficiency in comprehending intricate patterns and dependencies across both spatial and sequential dimensions. Its effectiveness

shines in tasks like text classification, where capturing local patterns and long-term dependencies holds paramount importance.

Here the Confusion Matrix and ROC Curve shown below for this study with Convolutional Neural Network (CNN) + Long Short-Term Memory (LSTM) + GlobalMaxPooling1D layers (Hybrid Model).

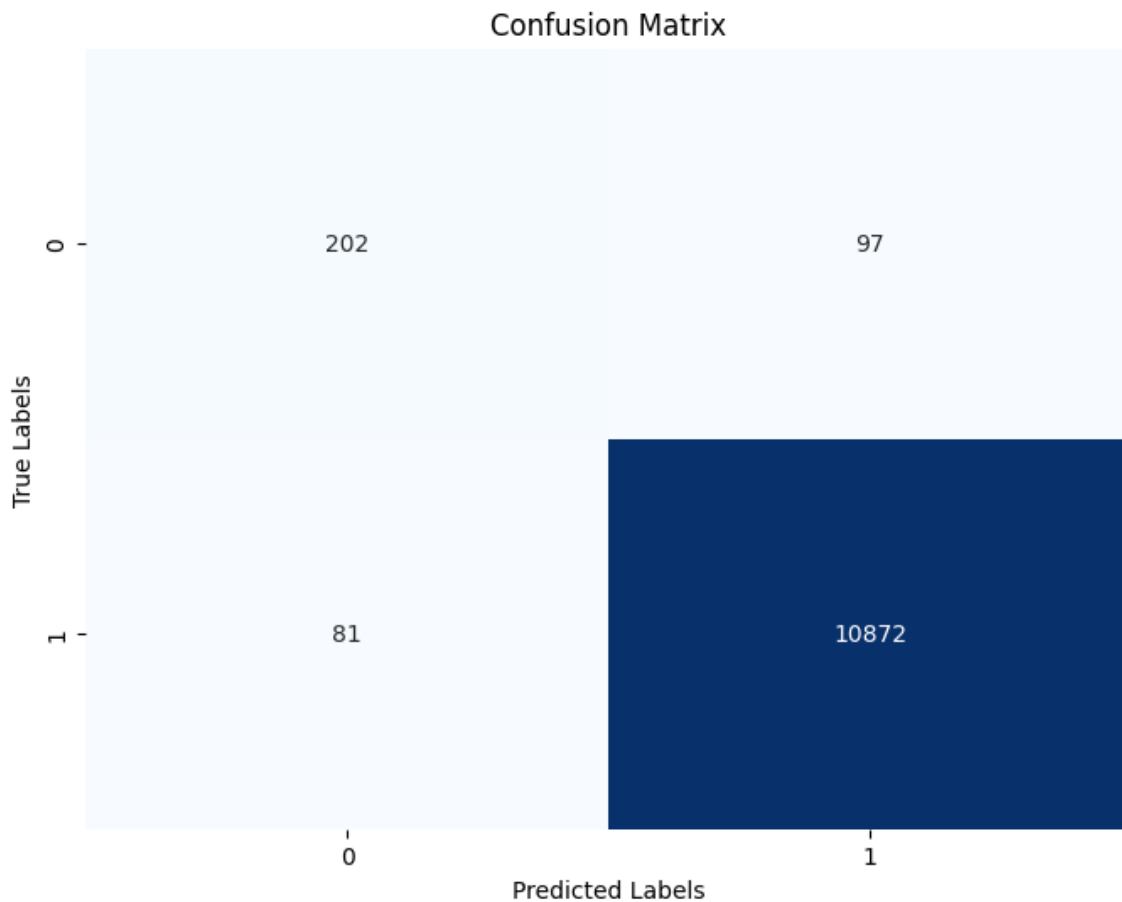


Fig 3.4.7 Confusion matrix

The Hybrid Model, incorporating Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and GlobalMaxPooling1D layers, has delivered outstanding results, attaining an accuracy of 98.41%, precision of 99.11%, recall of 99.26%, and an F1 score of 99.18%. This integrated approach underscores the model's effectiveness in precisely classifying data, maintaining a balanced performance in precision and recall, thereby showcasing its ability to differentiate between authentic and fake instances.

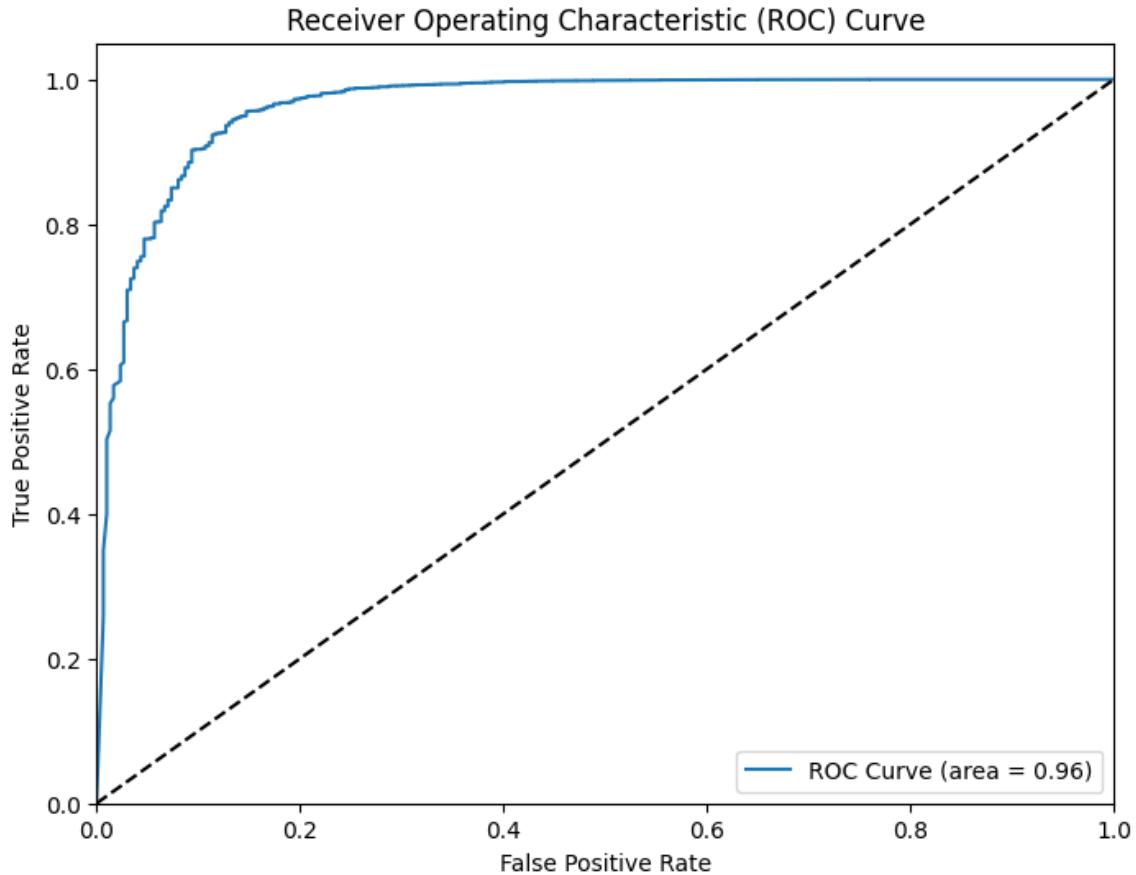


Fig 3.4.8 ROC Curve

3.5 Implementation Requirements

Bangla, being an inflectional language, often assigns multiple meanings to each word based on inflection. To detect and eliminate noise from each word, we employ various library functions that handle regular expressions, HTML tags, and punctuation. The implementation of this process requires the fulfillment of certain prerequisites.

Requirements:

1. **Python:** Python serves as the programming language for our model implementation, recognized as the optimal language for machine learning applications.
2. **Pandas:** Utilized as `pd` in our code, Pandas is a Python library for the manipulation and analysis of data.

3. **Numpy:** Numpy, another Python library, is employed for its high-performance multidimensional array capabilities and fundamental tools for array computation.
4. **Itertools:** Python's Intertools module provides a collection of utilities crafted for managing iterators, which are data varieties suitable for use in for loops.
5. **Matplotlib:** Matplotlib, a Python plotting library, along with its numerical extension NumPy, supplies an object-oriented API that allows the integration of plots into applications using general-purpose GUI toolkits such as wxPython, Tkinter, Qt, or GTK+.
6. **Sklearn:** Abbreviated as scikit-learn, this library is vital for learning and predicting. In scikit-learn, an estimator for classification is a Python object that implements the methods `predict(T)` and `fit(X, y)`. An instance of such an estimator is the class `sklearn.svm.SVC`, dedicated to support vector classification.

CHAPTER 4

Discussion and Experimental Results

4.1 Experimental Setup

To achieve optimal results, we recognize the necessity of a versatile model. The preparation of a high-performance computer is crucial for efficiently building the desired model. In our research, dealing with an extensive dataset posed some challenges, yet it also provided a significant opportunity to leverage Google Colab. Colab stands out as a popular and widely used research tool, facilitating swift dataset training. Despite its advantages, we encountered some noticeable challenges due to the large volume of data, resulting in slower processing. To address this, we employed the split method for model training, executing it with 20 epochs and a batch size of 512, patience 2.

Performance appraisal is crucial for comprehending the efficacy of a proficient classifier for making data-driven decisions. In our assessment of performance, we considered various metrics, including the scatterplot, recall, precision, accuracy, F1-score, and the AUC-ROC curve. Additionally, we presented the number of recommendations made correctly or incorrectly by the classifiers.

4.2 Experimental Results & Analysis

Comparison Between Accuracy Before and After Balancing Dataset: Evaluating the effectiveness of a machine learning model involves a critical examination of accuracy before and after implementing dataset balancing techniques. Balancing the dataset becomes essential when class distribution is imbalanced, signifying a substantial numerical difference between classes.

Before Balancing: Prior to implementing balancing techniques, the model undergoes training on an unadjusted dataset, where one class might exhibit dominance over the other. In such instances, the model tends to display bias towards the majority class, resulting in suboptimal performance, particularly for the minority class. The accuracy metric may appear deceptively high, as the model may predominantly predict the majority class while overlooking the minority class.

The accuracy table shown below of the algorithms or model used in this study:

TABLE 4.2.1: Before Balancing the Dataset

Models	Accuracy	Precision	Recall	F1 score
Bidirectional GRU (Proposed Model)	99.07%	99.10%	99.54%	99.52%
1D CNN	98.73%	98.75%	99.96%	99.35%
LSTM	98.20%	98.68%	99.47%	99.08%
CNN + LSTM + GlobalMaxPooling1D	98.55%	98.94%	99.57%	99.25%

After Balancing: Following the implementation of data balancing methods, which may involve oversampling the minority class, under sampling the majority class, or utilizing advanced techniques like SMOTE (Synthetic Minority Over-sampling Technique), the dataset undergoes modification to achieve a more balanced distribution between classes. The objective is to enhance the model's capacity to generalize effectively across both classes.

The accuracy table shown below of the algorithms or model used in this study:

TABLE 4.2.2: After Balancing the Dataset

Models	Accuracy	Precision	Recall	F1 score
Bidirectional GRU (Proposed Model)	99.13%	99.22%	99.89%	99.55%
1D CNN	98.91%	99.12%	99.77%	99.44%
LSTM	98.57%	99.15%	99.38%	98.27%
CNN + LSTM + GlobalMaxPooling1D	98.41%	99.11%	99.26%	99.18%

Errors are inherent in human cognition, whereas precision and authenticity distinguish machines from humans. Although our model yielded satisfactory results overall, there were exceptions. Meanwhile, Bidirectional GRU (Proposed Model) demonstrated exceptional results among the algorithms.

4.3 Discussion

The proliferation of digital media worldwide has elevated the significance of addressing the pervasive issue of fake news in society. Mitigating the influence of fake news in our daily lives is a formidable challenge. While identifying fake news poses a significant challenge for humans, machines are adept at discerning such misinformation. However, training machines with appropriate data is a complex task. If the data collected from various online sources contains numerous errors and noise, it can significantly impact the accuracy of predictions. To circumvent such challenges, we prioritized preprocessing our data meticulously, eliminating errors and addressing missing values. We conducted multiple executions of our model in both noisy and noise-free environments, yielding promising and encouraging results. The outcomes instill confidence in our work, suggesting its resilience and promising prospects for future advancements.

CHAPTER 5

Environment, Impact on Society and Sustainability

5.1 Impact on Society

Fake news is a big problem, and it's not just about individual harm. It's like a threat to the way our society works. Social media is a big part of our lives, and it makes fake news even more powerful. When people believe in fake stories, they might get into trouble without knowing it. That's why we need good ways to find and stop fake news.

My goal is to use automated systems to find fake news early on and stop it from causing problems. Getting rid of all fake news is really hard, but we want to help people tell what's true and what's not. By slowing down how fake news spreads, we hope to make a world where false information has less power, and people can be more sure about what's real. As technology gets better, finding ways to spot fake news becomes really important to keep information honest and protect our society.

5.2 Impact on Environment

In the intricate tapestry of our information ecosystem, the pervasive infiltration of misinformation, particularly through the channels of fake news, weaves a narrative that reverberates across the expanse of our environmental consciousness. Beyond its seemingly innocuous facade, this misinformation harbors the potential to not only skew public perceptions but also to imprint lasting imprints on the decisions that govern our environmental ethos.

Our dedicated efforts to stem the relentless tide of false information spring from a profound aspiration—to foster an enlightened society that is not just well-informed but equipped with the discernment needed to sculpt the path forward for environmental stewardship. As the custodians of factual integrity, we aspire to empower individuals with a robust knowledge base. This knowledge, wielded wisely, becomes a formidable force in steering decisions that resonate with the harmonious coexistence of humanity and the environment. In essence, our commitment is a testament to the belief that a well-informed society is not just a guardian of truth but also a custodian of the environmental legacy we bequeath to future generations. Through the vigilant dissemination of accurate information, we

endeavor to forge a collective consciousness capable of navigating the intricate landscapes of environmental policies and practices with sagacity and purpose.

5.3 Ethical Aspects

Furthermore, our unyielding commitment to advancing knowledge is not merely a source of inspiration but a guiding principle that propels us to actively contribute and elevate the existing research landscape. With a critical discernment of gaps in prior studies, our mission is to fortify the bedrock of investigative efforts, ensuring a robust foundation for future inquiry. The meticulous curation of additional data is a testament to our resolute dedication, intricately enhancing the dataset and broadening the vista of our study to unveil nuanced and comprehensive insights. Each layer of information meticulously gathered serves as a building block, fostering a more profound understanding of the subject matter and pushing the boundaries of scholarly exploration.

5.4 Sustainability Plan

In our unwavering pursuit of excellence, we meticulously orchestrated the integration of deep learning and machine learning classifiers, navigating to the expansive landscape of data intricacies. This deliberate fusion of methodologies represents a strategic marriage, where the nuanced strengths of machine learning harmonize with the intricate capabilities of deep learning, yielding a formidable alliance poised for groundbreaking results. Our commitment to academic rigor propelled us to embark on a transformative journey, where the expansion of our dataset unfolded as a carefully choreographed symphony. Each data point curated with precision served not only to augment the depth of our investigation but also to sculpt a narrative that resonates with the evolving challenges of our times.

As avid researchers, we conducted an exhaustive review of related studies, meticulously identifying gaps and nuances that became the focal points of our dedicated efforts. In addressing these gaps, we fortified the foundations of our investigation, ensuring that our work stands as a testament to scholarly excellence and innovation.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

Our study has been dedicated to the realm of Bengali Natural Language Processing (NLP), striving to make impactful contributions to the identification and mitigation of fake news in the Bangla language. Emphasizing a holistic approach, we employed various machine learning algorithms alongside an extensive dataset, fostering a robust foundation for our work. The rigorous two-month timeline of our research allowed us to navigate through multifaceted aspects. Our meticulously designed workflow, illustrated in Figure 6.1, served as a comprehensive guide in conducting our investigation step by step. This structured methodology facilitated a thorough exploration of the intricacies involved in detecting and addressing fake news in the Bengali language.

With the broader goal of advancing Bengali NLP exploration, our research has the potential to act as a catalyst for further developments in the field. Beyond merely identifying fake news, our work aims to lay the groundwork for programmable detectors tailored to the nuances of the Bengali language. As we delve into the intricate layers of linguistic subtleties, our research aspires to contribute meaningfully to the ongoing discourse surrounding fake news detection in the Bengali language.

6.2 Conclusions

In our research, I have amalgamated various techniques to construct a model designed for discerning the authenticity of news articles—distinguishing between veracity and falsehood. The swift dissemination of news in this digital age often leads to unforeseen circumstances, exacerbated by the prevalence of deceptive information.

In our exploration of Bengali Natural Language Processing (NLP), our research has delved into the intricate realm of fake news identification. Employing a combination of both machine learning as well as deep learning algorithms has yielded remarkable results. The extensive dataset, comprising over 60k data points sourced from diverse online media and news platforms, served as the foundation for our investigative endeavors. Our primary

objective was to discern the authenticity of news articles, distinguishing between genuine and fake narratives. Remarkably, all our machine learning algorithms exhibited an impressive accuracy but the most impressive accuracy from Bidirectional GRU Model accuracy rate of approximately 99.13%.

While our final results are promising, it is crucial to acknowledge the inherent challenges and imperfections that accompany such research endeavors. Despite the inherent odds and faults, our meticulous efforts have culminated in achieving our desired outcome – a robust model adept at detecting Bengali fake information's with remarkable ease.

6.3 Implication for Further Study

In our current implementation, we utilized three different models, each yielding distinct accuracy values. It's important to highlight that our focus was solely on models with favorable accuracy scores.

Fake news creates a threat not only to individuals but also to our societal fabric, akin to a cancer that distorts our judgment. Addressing this issue requires the proactive identification of fake news. Given the limited research on Bangla fake news detection, there is a compelling need for further exploration in this area. Challenges, particularly related to dataset quantity, underscore the importance of expanding our dataset. We aim to highlight the distinctions in fake news detection between English and Bangla languages. Looking ahead, our vision extends beyond model implementation. We aspire to develop a comprehensive website that empowers users to verify news sourced from various platforms— online news portals, social media, and other websites. This initiative aims to combat confusion arising from fake news. The envisioned website will feature user-friendly options for determining the authenticity of news articles. Users can conveniently navigate the platform, copy-paste news snippets, and promptly discern whether the information is true or false. This endeavor marks a pioneering effort, as there is currently a lack of dedicated Bangla Fake News detectors. By introducing this website, we aim to fill this gap, providing a reliable tool for users to verify the accuracy of news and ultimately reduce confusion stemming from misinformation.

REFERENCES

- [1] M. Z. HOSSAIN, M. A. RAHMAN, M. S. ISLAM AND S. KAR, "BANFAKENEWS: A DATASET FOR DETECTING FAKE NEWS IN BANGLA," ARXIV PREPRINT ARXIV:2004.08789, 19 APR 2020.
- [2] M. G. HUSSAIN, M. R. HASAN, M. RAHMAN, J. PROTIM AND S. A. HASAN, "DETECTION OF BANGLA FAKE NEWS USING MNB AND SVM CLASSIFIER," 2020 INTERNATIONAL CONFERENCE ON COMPUTING, ELECTRONICS & COMMUNICATIONS ENGINEERING (ICCECE), PP. 81-85, 2020.
- [3] S. T. UDDIN, M. R. SHAHRIAR, F. RIZON, R. A. POLOCK AND S. . I. SHAMEEM, "FAKEDTECT: BANGLA FAKE NEWS DETECTION MODEL BASED ON DIFFERENT MACHINE LEARNING CLASSIFIERS," DOCTORAL DISSERTATION, BRAC UNIVERSITY, 2021.
- [4] O. SHARIF, M. . M. HOQUE, A. S. M. KAYES, R. NOWROZY AND I. H. SARKER, "DETECTING SUSPICIOUS TEXTS USING MACHINE LEARNING TECHNIQUES," APPLIED SCIENCES, VOL. 10 , NO. 18, P. 6527, 2020.
- [5] M. Z. H. GEORGE, N. HOSSAIN, M. R. BHUIYAN, A. K. M. MASUM AND S. ABUJAR, " BANGLA FAKE NEWS DETECTION BASED ON MULTICHANNEL COMBINED CNN-LSTM," 2021 12TH INTERNATIONAL CONFERENCE ON COMPUTING COMMUNICATION AND NETWORKING TECHNOLOGIES (ICCCNT), PP. 1-5, JULY 2021.
- [6] A. BALO, J. ISLAM AND A. . A. BAKI , "BENGALI FAKE NEWS DETECTION USING MACHINE LEARNING," 2019.
- [7] 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," 2021 5TH INTERNATIONAL SYMPOSIUM ON MULTIDISCIPLINARY STUDIES AND INNOVATIVE TECHNOLOGIES (ISMSIT), PP. 442-447, 2021.
- [8] A. MAHABUB, "A ROBUST TECHNIQUE OF FAKE NEWS DETECTION USING ENSEMBLE VOTING CLASSIFIER AND COMPARISON WITH OTHER CLASSIFIERS," SN APPLIED SCIENCES VOLUME 2, PP. 1-9, 2020.
- [9] F. AKTER, S.. A. TUSHAR, S. A. SHAWAN, M. KEYA, S. A. KHUSHBU AND S. ISALM, "SENTIMENT FORECASTING METHOD ON APPROACH OF SUPERVISED LEARNING BY NEWS COMMENTS," 2021 12TH INTERNATIONAL CONFERENCE ON COMPUTING COMMUNICATION AND NETWORKING TECHNOLOGIES (ICCCNT), PP. 1-7, JULY 2021
- [10] KAI SHU, DEEPAK MAHUESWARAN, SUHANG WANG, DONGWON LEE, HUAN LIU, "FAKE NEWS NET: A DATA REPOSITORY WITH NEWS CONTENT, SOCIAL CONTEXT AND SPATIAL TEMPORAL INFORMATION FOR STUDYING FAKE NEWS ON SOCIAL MEDIA" THE 9TH INTERNATIONAL CONFERENCE ON EMERGING UBIQUITOUS SYSTEMS AND PERVASIVE NETWORKS (SUBMITTED ON 5 SEP 2018 (v1), LAST REVISED 27 MAR 2019 (THIS VERSION, v3).
- [11] KAI SHU, AMY SLIVA, SUHANG WANG, JILIANG TANG, HUAN LIU, " FAKE NEWS DETECTION ON SOCIAL MEDIA A DATA MINING PERSPECTIVE" ACM SIGKDD EXPLORATIONS NEWSLETTER, VOLUME 19 ISSUE 1, JUNE 2017.

- [12] NAMAN SINGH ; TUSHAR SHARMA ; ABHA THAKRAL ; TANUPRIYA CHOUDHURY “DETECTION OF FAKE PROFILE IN ONLINE SOCIAL NETWORKS USING MACHINE LEARNING” 2018 INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING AND COMMUNICATION ENGINEERING (ICACCE)
- [13] VERÓNICA PÉREZ-ROSAS, BENNETT KLEINBERG, ALEXANDRA LEFEVRE, RADAMIHALCEA, “AUTOMATIC DETECTION OF FAKE NEWS”, COMPUTATION AND LANGUAGE (CS.CL) SUBMITTED ON 23 AUG 2017.
- [14] BENJAMIN RIEDEL, ISABELLE AUGENSTEIN, GEORGIOS P. SPITHOURAKIS, SEBASTIAN RIEDEL, “A SIMPLE BUT THOUGHT-BEAT BASELINE FOR THE FAKE NEWS CHALLENGE STANCE DETECTION TASK” (SUBMITTED ON 11 JUL 2017 (v1), LAST REVISED 21 MAY 2018 (THIS VERSION, v2)).
- [15] HADEER AHMED, AUTHORISS, TRAORE SHERIF SAAD, “DETECTION OF ONLINE FAKE NEWS USING N-GRAM ANALYSIS AND MACHINE LEARNING TECHNIQUES, CONFERENCE PAPER, FIRST ONLINE: 11 OCTOBER 2017.
- [16] MINYOUNG HUH, ANDREW LIU, ANDREW OWENS, ALEXEI A. EFROS, “FIGHTING FAKE NEWS IMAGE SPLICE DETECTION VIA LEARNED SELF-CONSISTENCY” COMPUTER VISION AND PATTERN RECOGNITION (CS.CV).
- [17] BENJAMIN D. HORNE, SIBEL ADALI, “FAKE NEWS PACKS A LOT IN TITLE, USES SIMPLER, REPETITIVE CONTENT IN TEXT BODY, MORE SIMILAR TO SATIRE THAN REAL NEWS” PUBLISHED AT THE 2ND INTERNATIONAL WORKSHOP ON NEWS AND PUBLIC OPINION AT ICWSM.
- [18] EMERSON F. CARDOSO RENATO, M. SILVA TIAGO, A. ALMEIDA, “TOWARDS AUTOMATIC FILTERING OF FAKE REVIEWS” VOLUME 309, 2 OCTOBER 2018.
- [19] W. KEN REDEKOP, “FAKE NEWS, BIG DATA, AND THE OPPORTUNITIES AND THREATS OF TARGETED ACTIONS” HEALTH POLICY AND TECHNOLOGY, 7(2), 113-114, 2018.
- [20] VERONICA PÉREZ-ROSAS¹, BENNETT KLEINBERG², ALEXANDRA LEFEVRE¹ RADA MIHALCEA¹, “AUTOMATIC DETECTION OF FAKE NEWS” UNIVERSITY OF MICHIGAN² DEPARTMENT OF PSYCHOLOGY, UNIVERSITY OF AMSTERDAM.
- [21] A. PETERS, E. TARTARI, N. LOTFINEJ, P. PARNEIX, D. PITTET, “FIGHTING THE GOOD FIGHT: THE FALLOUT OF FAKE NEWS IN INFECTION PREVENTION AND WHY CONTEXT MATTERS” 2018 PUBLISHED BY ELSEVIER LTD ON BEHALF OF THE HEALTHCARE INFECTION SOCIETY.
- [22] S. MOJANG PHD, TIEMING GENG, JO-YUN QUEENIE, LIARUOFAN XIA, CHIN-TSER HUANG PHD, HWALBIN KIM PHD, JIJUN TANG PHD, “A COMPUTATIONAL APPROACH FOR EXAMINING THE ROOTS AND SPREADING PATTERNS OF FAKE NEWS: EVOLUTION TREE ANALYSIS” 2018 ELSEVIER LTD.
- [23] MAURIDHI HERY PURNOMO, SURYA SUMPENO, ESTHER IRAWATI SETIAWAN, DIANA PURWITASARIA, “KEYNOTE SPEAKER II: BIOMEDICAL ENGINEERING RESEARCH IN THE SOCIAL NETWORK” ANALYSIS ERA: STANCE CLASSIFICATION FOR ANALYSIS OF HOAX MEDICAL NEWS IN SOCIAL MEDIA”, 2017 PUBLISHED BY ELSEVIER B.
- [24] MONTHER ALDWAIRI, ALI ALWAHEDI, “DETECTING FAKE NEWS IN SOCIAL MEDIA NETWORKS” VOLUME 141, 2018, PAGES 215-222.

- [25] AVARO FIGUEIRA, LUCIANA OLIVEIRA, "THE CURRENT STATE OF FAKE NEWS: CHALLENGES AND OPPORTUNITIES" CENTERIS / PROJMAN / HCIST 2017, 8-10 NOVEMBER 2017, BARCELONA, SPAIN VOLUME 121, PAGES 817-825.
- [26] SHOLK GILDA "EVALUATING MACHINE LEARNING ALGORITHMS FOR FAKE NEWS DETECTION" 2017 IEEE 15TH STUDENT CONFERENCE ON RESEARCH AND DEVELOPMENT (SCORED).

BANGLA FAKE NEWS IDENTIFICATION USING DEEP LEARNING AND MACHINE LEARNING

ORIGINALITY REPORT

8%

SIMILARITY INDEX

7%

INTERNET SOURCES

4%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Daffodil International University

Student Paper

2%

2

www.mdpi.com

Internet Source

1%

3

www.researchgate.net

Internet Source

1%

4

Submitted to Morgan Park High School

Student Paper

1%

5

dspace.daffodilvarsity.edu.bd:8080

Internet Source

1%

6

Risul Islam Rasel, Anower Hossen Zihad, Nasrin Sultana, Mohammed Moshiul Hoque. "Bangla Fake News Detection using Machine Learning, Deep Learning and Transformer Models", 2022 25th International Conference on Computer and Information Technology (ICIT), 2022

Publication

1%

getallcourses.net