Research Paper

Combating Deepfake News in Bengali: Detection and Prevention of Manipulated Social Media's Content Using Machine Learning

RESEARCH SUPERVISOR

Dr. Md. Ashikur Rahman Khan
Professor & Chairman
Department of ICE, NSTU

SUBMITTED BY

Md. Jane Alam
Roll No: ASH2111MSc112M
Session: 2020-2021
Year II, Term I
Department of ICE, NSTU

DEPARTMENT OF INFORMATION AND COMMUNICATION ENGINEERING

NOAKHLAI SCIENCE AND TECHNOLOGY UNIVERSITY

MARCH 2024

Combating Deepfake News in Bengali: Detection and Prevention of Manipulated Social Media's Content Using Machine Learning

MD. JANE ALAM

This research paper submitted to the Department of Information and Communication Engineering, Noakhali Science and Technology University, in partial fulfillment of the requirements for the degree of M.Sc. (Eng.) in Information and Communication Engineering.

DEPARTMENT OF INFORMATION AND COMMUNICATION ENGINEERING

NOAKHLAI SCIENCE AND TECHNOLOGY UNIVERSITY

MARCH 2024

DECLARATION

I hereby declare that this research paper has not been submitted elsewhere for the requirement of any kind of Degree, Diploma or Publication.

Md. Jane Alam
Roll No: ASH2111MSc112M
Session: 2020-2021
Year II, Term I
Department of Information and Communication Engineering
Noakhali Science & Technology University
Noakhali - 3814.

ACCEPTANCE

This research paper is submitted to the Department of Information & Communication Engineering, Noakhali Science & Technology University, Noakhali - 3814, Noakhali in partial fulfillment of the requirements for having the M.Sc. (Eng.) degree in ICE.

Dr. Md. Ashikur Rahman Khan
Professor & Chairman
Department of Information and Communication Engineering
Noakhali Science & Technology University
Noakhali - 3814.

TO MY BELOVED PARENTS

ABSTRACT

With the rapid growth of digital media and social networks, the spread of misinformation and manipulated content has become a significant concern. Nowadays, misinformation and false rumors mostly come from social media platforms. Deepfake technology, which allows the creation of realistic but fabricated news, audio, images, and videos, has emerged as a potent tool for the production and dissemination of deceptive content. While the detrimental impacts of deepfake news are widely recognized, research efforts in combating this phenomenon are largely focused on English and a few other major languages. On the other hand, Bengali grammar and language are significantly more difficult and essential to learn. This research paper aims to address this gap by proposing a comprehensive approach for detecting and preventing deepfake news specifically in the Bengali language.

This study utilizes a newly released set of Bangla fake news dataset, marked by experts. It then employs Bengali-based embeddings for machine learning classifiers and utilizes trained bidirectional encoder representations from transformers (BERT) in Bengali to determine sentiments in Bengali grammar.

The results show that increasing the training data consistently enhanced the BERT-based classifiers' performance, surpassing that of ML classifiers. When compared to prior studies on detecting Bengali fake news, the current study's findings indicate a more significant improvement.

Keyword: Bangla Fake News, BERT classifier, Fake News Detection, Bangla

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 01

INTRODUCTION

1.1 Introduction

Fake news is information that has been intentionally created to deceive or manipulate audiences, sometimes presented as authentic news but that is actually fake or misleading. It can involve fabricated stories, manipulated facts, or distorted interpretations presented as legitimate news. The aim of fake news is typically to sway opinions, push certain agendas, or cause confusion among the audience. The main aim of detecting fake news is to help online users steer clear of different kinds of false information.

In Bangladesh, social media and online news sources have grown in popularity over the last few years. Furthermore, accessing the news portals and the associated social media pages is not too difficult. These two platforms exchange every aspect of news from across the country. For all types of internet users, news portals are an essential component of online communication. Press releases, publications, articles, blogs, and other news-related material can be shared on these channels [1]. With busy schedules, many can't read newspapers to catch up on yesterday's news. So, they turn to websites and electronic media for updates. They offer a quick way for individuals to access the latest happenings. Whether it's news articles, stories, or data on these platforms, their influence is substantial. The content shared here significantly shapes how people view and comprehend current events, making these digital sources pivotal in molding public opinion and awareness [2]. The data we take in significantly influences our decision-making processes and shapes our worldview [3].

Because of this, fake news is extremely dangerous. The main disadvantage of online news portals and various social media platform are the spread of fake news, even though they have made our lives easier by providing access to daily news from anywhere. Fake news can be created for various reasons. Sometimes, it's for political agendas or to influence public opinion. At other times, it

might be for monetary gain through increased website traffic or ad revenue. Misinformation might also stem from misunderstandings or misinterpretations that spread rapidly before being corrected. In some cases, it's intentional manipulation to deceive or create chaos within society. Social media fake news has the power to destroy not only our society but the entire nation in a matter of hours, spreading like wildfire [4].

The global impact of false information is causing widespread chaos. In the 2016 US election, a quarter of the American population reportedly visited deceptive news websites over a six-week period, potentially influencing the election's outcome [5]. Similarly, in Bangladesh, the 2012 Ramu incident stands out as a poignant example. Nearly 25,000 individuals participated in the destruction of Buddhist temples triggered by a fake Facebook post from an anonymous account [6]. The incident resulted in the destruction of approximately 12 Buddhist temples, several monasteries, and around 50 houses due to the collective rage fueled by the false information. Instances involving fake news containing sacrilegious content have the potential to incite such emotional reactions in communities where religious sentiments run high, leading to similar regrettable events.

In the effort to combat misinformation, specialized platforms such as www.snopes.com, www.fullfact.org, www.jaachai.com, www.politifact.com, and www.factcheck.org have emerged. These sites manually curate and debunk potentially false news stories from online sources, providing logical and factual explanations behind their inaccuracy. However, these platforms face limitations in their responsiveness to rapidly emerging fake news events. Recently, computational methods have been introduced as an additional strategy in the fight against fake news, aiming to supplement these manual efforts for quicker and more adaptive responses to combat the spread of misinformation. Long [7] explored using multi-perspective speaker profiles to spot fake news, while Yang [8] employed linguistic cues to detect satirical news. Karadzhov [9] proposed an automated fact-checking model that cross-checks news claims with external sources. In the realm of social media, Dong [10] utilized deep two-path semi-supervised learning for fake news detection, Alawadh [11] used Mini-BERT Technology to detect and analyze Arabic Fake News. Few studies have examined the existence of fake news before the common era (BCE), but the concept reemerged with the advent of printed media [12]. However, among the transformative impacts of news-disseminating platforms, the emergence of social media stands out as the primary

catalyst for the proliferation of fake news. Social media platforms drastically amplify the spread of news [13]. Fake news fabricates information regarding individuals, objects, or events, causing uncertainty among users about its accuracy. In certain cases, it can lead to mental health issues, while deceptive information about companies can significantly harm their reputation and business prospects [14].

Because of the increasing difficulty in automatically identifying fake news, recent research has concentrated on addressing these concerns. There's been a recent surge in interest surrounding social media and other news content [15]. Additionally, countries dealing with fake news challenges are witnessing a rise in the utilization of Bangla news channels and social media platforms.

However, these studies focused solely on analyzing English or other language news content. Despite Bangla being spoken by around 300 million people worldwide, ranking as the 7th most spoken language globally, there isn't a computational method or resource available right now to address the problem of fake news in Bengali. The major challenge lies in assembling a authentic and balanced dataset in real time, incorporating suitable attributes. Limited research has employed both private and public datasets. Similarly, our proposed study utilizes a newly released public Bangla news dataset derived from real-time data encompassing both reliable and unreliable news sources. Studies into fake news have found misleading cues, informal news styles, and fabricated sentiments as prominent traits.

In the past, fake news detection relied on sentiments, linguistic hints, and writing style. However, the explosion of deep learning models in discourse and sentiment analysis drove us to create a DL-focused solution. Our proposed study makes the following contributions towards this endeavor:

- Making Specialized Bengali Datasets: Creating customized datasets in Bengali with varied content to train machine learning models that spot deepfake news in the language.
- Integration of BERT Technology for Enhanced Bengali Deepfake Detection: Leveraging state-of-the-art BERT (Bidirectional Encoder Representations from Transformers) models to refine and bolster the accuracy of Bengali deepfake detection algorithms, marking a

significant advancement in combatting manipulated content within the realm of Bengali social media.

- Constructing robust machine learning algorithms customized for Bengali language nuances and socio-cultural contexts, aimed at effectively identifying and preventing the spread of manipulated content in social media platforms.
- Providing guidelines for integrating the deepfake detection & prevention model into existing media platforms to combat the spread of deepfake news effectively.

We expect our research work will play an important role in the development of fake news detection and prevention systems. In the rest of the paper, we provide a brief description of our dataset categorization techniques as well as the creation and effectiveness of systems for preventing and detecting fake news.

1.2 Problem Statement

Deepfake technology has developed as a major issue in the digital era, permitting the development of very convincing fake multimedia material that may confuse and misinform the public. While there has been substantial study on identifying deepfake material in English and other languages, there is a noteworthy lack of studies addressing the detection and mitigation of deepfake news in Bengali.

The absence of effective tools and techniques to identify manipulated multimedia content in Bengali poses a serious threat to the integrity of information and the potential for widespread dissemination of misleading or false information. As a result, there is an urgent need to develop robust machine learning algorithms and frameworks tailored for the detection of deepfake news in the Bengali language.

This research aims to address the gap by proposing and implementing novel approaches for detecting manipulated multimedia content, particularly deepfake videos and images, in Bengali. The study will explore the unique linguistic and cultural aspects of the Bengali language to develop language-specific models that can accurately identify deepfake content. Additionally, the research

will investigate the performance of various machine learning algorithms, such as deep neural networks and convolutional neural networks, in detecting deepfakes in the Bengali language.

The outcomes of this study will lead to the construction of a complete system capable of preventing deepfake news in Bengali, therefore protecting the reliability and authenticity of digital media material. Furthermore, the suggested methodologies may be incorporated into current media verification frameworks and platforms to improve their efficiency in detecting and combating deepfake material in Bengali.

1.3 Research Gap

The research gap in combating deepfake news in Bengali and detecting manipulated multimedia content using machine learning can be identified in the following areas:

a) **Limited Research on Bengali Language:** Most existing research on deepfake detection and combating fake news focuses on English language content. However, Bengali is a widely spoken language with its unique linguistic characteristics, cultural references, and contextual nuances [1]. The lack of dedicated research on deepfake detection in Bengali poses a significant gap, as the techniques and models developed for English may not be directly applicable to Bengali. Therefore, there is a need for research that specifically addresses the challenges and intricacies of detecting deepfake news in the Bengali language.

b) **Scarcity of Labelled Datasets:** Deepfake detection models rely on large-scale, labelled datasets for training and evaluation. However, there is a lack of comprehensive and publicly available labelled datasets specific to Bengali deepfake content [2]. The absence of such datasets hinders the development and evaluation of deepfake detection models tailored for the Bengali language. Closing this research gap requires the creation of curated datasets containing authentic and manipulated Bengali multimedia content, enabling researchers to train and validate models effectively.

c) **Contextual Understanding and Cultural Nuances:** Deepfake news often exploits cultural, social, and political contexts to manipulate public opinion [3]. Understanding the

cultural nuances and context-specific features of Bengali media is crucial for accurate deepfake detection. However, the existing research does not adequately explore the cultural aspects and contextual understanding necessary for detecting manipulated multimedia content in Bengali. Bridging this gap involves studying the specific cultural references, language usage, and media consumption patterns in the Bengali-speaking population to develop effective detection methods.

By addressing these research gaps, the proposed research project can contribute significantly to the development of effective deepfake detection methods and frameworks tailored for Bengali, enabling the detection and mitigation of manipulated multimedia content in the Bengali media landscape.

1.4 Relevance to National Interest

Combating deepfake news in Bengali and detecting manipulated multimedia content using machine learning is of significant relevance to national interest due to the following reasons:

a) **Preserving Information Integrity:** Deepfake news poses a severe threat to the integrity of information and the credibility of media sources. By spreading fabricated and manipulated content, it can distort public perception, mislead citizens, and erode trust in news sources. Protecting the integrity of information is crucial for maintaining a well-informed society, promoting democratic values, and ensuring the stability of the nation.

b) **Safeguarding Social Cohesion:** Deepfake news has the potential to incite social unrest, amplify existing divisions, and polarize communities. By spreading misinformation, it can manipulate public opinion, fuel conflicts, and undermine social harmony. Combatting deepfake news in Bengali is essential for safeguarding social cohesion, fostering a sense of unity, and promoting healthy public discourse within the country.

c) **Protecting Democratic Processes:** Deepfake news can pose a significant threat to democratic processes, including elections and public decision-making. Manipulated multimedia content can be used to influence voter behaviour, discredit political opponents, and undermine the fairness and transparency of elections. Detecting and countering deepfake news in Bengali is vital for ensuring the integrity of democratic processes,

preserving the citizens' right to access accurate information, and upholding the principles of fair and free elections.

d) **Mitigating Security Risks:** Deepfake technology can be employed for malicious purposes, such as spreading propaganda, creating fake identities, or conducting targeted disinformation campaigns. These activities can have severe security implications, including compromising national security, inciting violence, or damaging the reputation of institutions and individuals. By developing robust deepfake detection mechanisms in Bengali, potential security risks can be mitigated, thereby protecting the nation's interests.

e) **Promoting Media Literacy:** The fight against deepfake news necessitates the promotion of media literacy and critical thinking skills among the public. By raising awareness about deepfake technology, its potential impact, and the techniques to identify manipulated content, individuals can make informed judgments and become less susceptible to misinformation. This, in turn, strengthens the overall media ecosystem, empowers citizens, and enhances the nation's resilience against disinformation campaigns.

f) **Strengthening Digital Infrastructure:** Addressing the challenges of deepfake detection in Bengali requires the development of robust machine learning models, infrastructure, and technical capabilities. Investing in research and innovation in this area not only enhances the nation's technological capacity but also strengthens the overall digital infrastructure. Such advancements can have broader applications beyond deepfake detection and contribute to the growth of the digital economy and technological competitiveness.

In summary, combating deepfake news in Bengali and detecting manipulated multimedia content using machine learning aligns with national interest by preserving information integrity, safeguarding social cohesion, protecting democratic processes, mitigating security risks, promoting media literacy, and strengthening the digital infrastructure. By addressing these issues, the proposed research project can contribute to a more informed, resilient, and secure society, thus serving the long-term interests of the nation.

1.4 Objectives

The objectives of this research are as follows:

i. To investigate and adapt existing machine learning algorithms and techniques for deepfake detection to the Bengali language.
ii. To design and implement a deepfake detection model specific to Bengali multimedia content.
iii. To evaluate the performance of the proposed model using benchmark datasets and real-world scenarios.
iv. To provide guidelines for integrating the deepfake detection model into existing media platforms to combat the spread of deepfake news effectively.

Contribution

The objectives of this research are as follows:

i. To investigate and adapt existing machine learning algorithms and techniques for deepfake detection to the Bengali language.
ii. To design and implement a deepfake detection model specific to Bengali multimedia content.
iii. To evaluate the performance of the proposed model using benchmark datasets and real-world scenarios.
iv. To provide guidelines for integrating the deepfake detection model into existing media platforms to combat the spread of deepfake news effectively.

Scope of the Research Work

CHAPTER 02

LITERATURE REVIEW

2.1 Related Works

This chapter offers a comprehensive analysis of the research on the use of various machine learning approaches to identify and prevent deepfake news." Certain limits and obstacles in the field have been recognized by the research that have been evaluated. Let us now look at the important study done by various researchers in this field.

2.2 Existing Fake News Detection Techniques

During the COVID-19 pandemic, a supervised machine learning system was created to identify false information and rumours on Arabic-language social media [16]. Using Twitter's Streaming API, the author gathered one million Arabic tweets and analysed them to identify pandemic-related topics, detect rumours, and predict tweet sources. A sample of 2,000 tweets was manually labelled as false information, correct information, or unrelated. With the use of many machine learning classifiers, including Naive Bayes, Logistic Regression, and Support Vector Machine, rumour identification was accomplished with 84% accuracy. However, this research has limitations, including dataset unavailability and reliance on a single source of rumours, the Saudi Arabian Ministry of Health.

Previously, AI-driven automated solutions proposed the use of machine learning (ML) and deep learning (DL) methods to identify fake news or text. The Arabic language dataset was collected using a variety of web-scraping techniques or from publicly available datasets, with noteworthy results. Using seven different ML classifiers, sarcasm detection in two Arabic news datasets (misogyny and Abu-Farah) demonstrated encouraging results, especially with the BERT technique. The binary classification achieved a 91% accuracy rate, while multiclass problems achieved an 89% accuracy using the same misogyny dataset. Furthermore, employing the BERT method for sarcasm detection demonstrated an 88% accuracy in binary classification and 77% accuracy in multiclass classification, surpassing previous methods [17].

A comparison between news websites involved ranking them against verified news sources to create a new accuracy score. This score was utilized to authenticate sites based on specific criteria. Assessing similarity between news items used the top1 cosine method, employing TF-IDF

analysis, and computing a 50% score through the top1 cosine method. An innovative accuracy measure, previously unexplored, was introduced and compared with prior studies to showcase its novel inclusion [18].

A research study utilized a gradient boosting technique to identify rumors within Arabic tweets, considering rumors as a form of extensive misinformation on social media platforms. Twitter data encompassing both rumors and non-rumors underwent preprocessing, extracting content-, user-, and topic-related features. Applying classical ML methods to the refined dataset, the eXtreme Gradient Boost (XG-Boost) achieved a high accuracy rate of 97.18% in classifying rumor-related data within 60% of the training dataset [19].

Perez-Rosas utilized a linear SVM classifier and linguistic features like N-grams, punctuations, and LIWC for supervised learning models to detect fake news patterns [20]. They collected a dataset of 240 genuine news articles from various US mainstream news sites and created a corresponding dataset with fake versions of these articles using crowdsourcing through Amazon Mechanical Turk (AMT). However, to uncover deeper features of fake news, neural networks, particularly word embedding techniques with deep neural networks, have proven effective [21]. Nonetheless, neural network models require extensive datasets. For their study, the Liar dataset—which included 12.8K human-labeled statements—was utilized, and deep neural network structures as well as surface-level linguistic analysis were applied [22].

Within a big data architecture, true and fake news were identified from political, governmental, US, Middle Eastern, and general news datasets. Following preprocessing and cleaning, contextual and statistical features were taken out and put into four different machine learning classifiers: Random Forest, Naïve Bayesian, Decision Tree, and SVM models. Furthermore, BERT, LSTM, and GRU techniques were employed, with a novel stacked model combining LSTM and GRU demonstrating superior performance. The findings of earlier investigations were surpassed by this stacked model, which obtained the best accuracies of 0.991 and 0.988 for the GRU and LSTM models, respectively [23].

A method called the soft voting classifier, utilized in studies like [24], combined classical machine learning classifiers such as Support Vector Machine, Logistic Regression, and Naïve Bayesian. These algorithms were trained and optimized using the FridSearchCV technique. Kaggle provided a dataset comprising both fake and real news for evaluation. With an accuracy rate of 93% and an F1 score with 94% precision and 92% recall, the ensemble technique was the most successful.

Various methods are utilized to assess fake news based on the specific news type being processed in [25]. This paper incorporates LIWC software for psycholinguistic analysis and utilizes Google's API to gauge news article toxicity, while TextBlob's API computes subjectivity. Different classifiers, including KNN, Naive Bayes, Random Forests, Support Vector Machine with RBF kernel, and XGBoost, are employed, assessed based on their AUC and F1 score performances. Notably, Naive Bayes showed subpar performance in both AUC (72%) and F1 score (75%). Conversely, Random Forests and XGBoost showcased stronger predictive abilities, achieving 86% and 85% accuracy, respectively, concerning AUC. Despite their proficiency, both models still have room for enhancement, possibly through newer classifier techniques. Acknowledging the limitations of sole textual analysis, the paper suggests leveraging deep learning and neural networks to trace news sources, a pivotal step in refining authenticity prediction.

The authors described a semi-supervised, graph-based approach for identifying bogus news in their publication [26]. They mainly used content-based detection algorithms, concentrating on three main components: creating an article similarity network, embedding articles into a Euclidean space, and applying graph learning techniques to infer missing labels. Notable innovations included identifying contextual commonalities between articles using a graph-based representation method and using word embeddings to create latent representations of news items in a lower-dimensional Euclidean space. Using Graph Neural Network topologies that are good at functioning with little labeled data, this system turned challenges involving the detection of fake news into a semi-supervised graph learning challenge. Comparable outcomes were found between 20% and 84% labeled data in their studies with different levels of labeled data, ranging from 2% to 20%. When analyzing fresh articles, their AGNN and GCN models fared better than others.

Ahmed et al. [27] detailed a methodology aimed at discerning pseudo-content, commencing with dataset preprocessing to remove stop words, punctuation, and extraneous characters. Subsequently, the text underwent tokenization and stemming to extract n-gram features, facilitating the formation of requisite documents. The final phase encompassed classifier training. Evaluation encompassed testing across six distinct machine learning algorithms, namely SGD, SVM, LSVM, K-Nearest Neighbor (KNN), LR, and DT. Veronica Perez-Rosas and associates looked into cutting-edge methods for spotting fake news online, with a particular emphasis on social media platforms, as reported in Pérez-Rosas et al. [28]. They carried out tests to produce accurate detectors and created two datasets specifically designed for the detection of fake news across several news domains. Utilizing a combination of lexical, syntactic, and semantic factors in addition to text readability attributes, they were able to detect bogus content with up to 78% accuracy.

In Monteiro et al. [29], linguistic analysis and machine learning were used to examine the detection of fake news in Brazilian Portuguese and other languages. Using text truncation and the Linear SVC approach, they produced an artificial classifier that achieved 89% accuracy. This allowed them to create the "Fake.Br corpus" by gathering and classifying authentic news items.

In the meanwhile, to strengthen AI methods against disinformation, Hanselowski et al. [30] performed a retrospective analysis on the top three systems in a false news challenge. The work offered thorough explanations of these systems, assessed the experimental design critically, suggested a new metric that included F1, and examined the features that were used.
A method for automatically categorizing tweets on Twitter into three different categories—hateful, offensive, and clean—was proposed by Gaydhani and colleagues (Gaydhani et al., [31]). They experimented with n-grams as features and their term frequency-inverse document frequency (TFIDF) values in several machine learning models, using a Twitter dataset. They examined various values of n in n-grams, compared the models, and used TFIDF normalizing approaches.

2.4 Summery:

Table 1 displays an overview of these investigations.

Table 1: Summery of the investigations

Title	Methods	Results	Data Source
Artificial intelligence-based approach for misogyny and sarcasm detection from Arabic texts [17]	BERT and other classical machine learning methods	Misogyny: Binary = 91%, Multi = 89% Sarcasm: Binary = 88%, Multi = 77%	Misogyny and Abu Farah datasets
An intelligent cybersecurity system for detecting fake news in social media websites [18]	Website ranking + Tf-IDF scores used to calculate the proposed news' accuracy	50% score through the top1 cosine method	Several benchmark datasets generated by Kaggle
An effective approach for rumor detection of Arabic tweets using extreme gradient boosting method [19]	Content-user, and topic-based features with machine learning classifiers	Accuracy using XG-boost = 97.18%	Public dataset
Context-Based Fake News Detection Model Relying on Deep Learning Models [23].	BERT, LSTM, and GRU techniques	Highest Accuracies of GRU: 0.991 and LSTM: 0.988	Information Security and Object Technology (ISOT) dataset
An Ensemble Machine Learning Approach for Fake News Detection and Classification Using a Soft Voting Classifier [24].	FridSearchCV technique	Accuracy = 93%, F1 Score = 93%, Precision = 94% and Recall = 92%	Kaggle dataset
A hybrid deep model for fake news detection," in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management [25]	KNN, Naive Bayes, Random Forests, SVM with RBF kernel, and XGBoost	Naive Bayes: AUC = 72%, F1 Score = 75%, Random Forests: 86%, XGBoost: 85%	Two real world datasets

Throughout our literature review, we have found that most research studies focus on fake news detection methods in English and other languages. However, these methods have difficulties in

finding and using an authentic and well-balanced dataset along with matching ground-truth annotations. Since research on fake news in Bengali language is still in its early stages, so we create a method to identify fake news in Bangla and adapt our dataset in a variety of ways that may be applied to other research streams.

SUMMARY

CHAPTER 03

METHODOLOGY

3.1 Methodology

The main goal of this research work was to detect and prevent manipulated social media deepfake news specifically in the Bengali language. This goal has been achieved by rigorously evaluating and comparing the performance of multiple machines learning methods, including Logistic Regression (LR), Decision Tree (DT), Gradient Boosting (GB), Random Forests (RF) and Bidirectional Encoder Representations from Transformers (BERT), and various performance metrics. Ultimately, this research project aims to bridge the research gap by developing innovative approaches and models for detecting deepfake news specifically in Bengali.

3.2 Machine Learning Techniques / Algorithms

In our research we've used various machine learning models that are suitable for deepfake detection in Bengali. This includes traditional machine learning algorithms, deep learning architectures (e.g.,

i. Logistic Regression (LR)
ii. Random Forest Classification (RF)
iii. Decision Tree (DT)
iv. Gradient Boosting (GB)
v. Bidirectional Encoder Representations from Transformers (BERT)

or state-of-the-art models specifically designed for deepfake detection. Train the models using the annotated dataset, considering appropriate data augmentation techniques to enhance model performance and generalization.

3.2.1 Logistic regression

Logistic regression, a supervised machine learning technique, is predominantly employed for classification tasks aiming to forecast the probability of an instance belonging to a specific class. This statistical method assesses the relationship between a group of independent variables and a set of binary dependent variables. It serves as a valuable tool for decision-making, utilizing a set of independent factors to predict the categorical dependent variable. A discrete or categorical value is required as the result of a categorical dependent variable's prediction using logistic regression. It provides probabilistic values between 0 and 1, as opposed to exact values between 0 and 1. This can appear as True or False, 0 or 1, Yes or No, and so on. Rather of fitting a regression line to 0 or 1, logistic regression uses a sigmoid-shaped logistic function to predict two maximum values. The logistic function's curve shows the likelihood of an event dependent on its weight. Logistic regression is a strong machine learning technique that may be used to categorize new data and provide probabilities based on both continuous and discrete datasets.

a) **Terminologies involved in Logistic Regression:** The following are some terminology frequently used in logistic regression:
   - **Independent variables:** Also referred to as predictor variables or features, these are the variables used to predict the dependent variable. They can be continuous, categorical, or binary.
   - **Dependent variable:** Also known as the response variable or target variable, it is the variable being predicted in logistic regression. It is typically categorical or binary.
   - **Logistic function:** Also called the sigmoid function, it is used to transform the linear combination of independent variables and coefficients into probabilities between 0 and 1.
   - **Logit:** The natural logarithm of the odds ratio. In logistic regression, the relationship between the independent variables and the dependent variable is modelled using the logit function.
   - **Odd ratio:** The ratio of the probability of an event occurring to the probability of it not occurring. In logistic regression, the coefficients represent the change in the odds of the dependent variable given a one-unit change in the independent variable.

- **Log-odds:** The natural logarithm of the chances is called the log-odds, or logit function. The log odds of the dependent variable in logistic regression are represented as a linear mixture of the intercept and the independent factors.

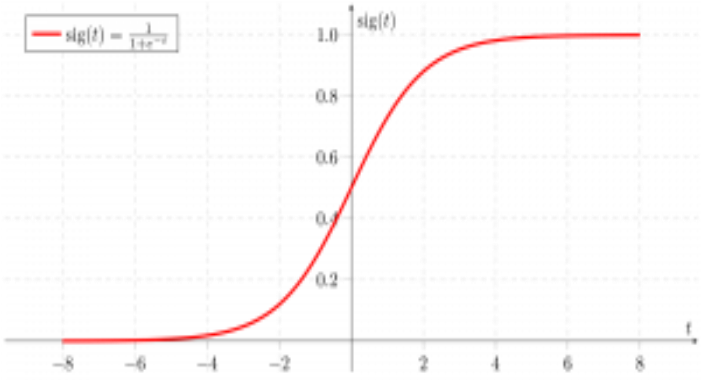

Figure 3.1: Logistic function (sigmoid function)

- **Coefficient:** Also known as beta coefficients or weights, these represent the strength and direction of the relationship between each independent variable and the dependent variable.
- **Intercept:** The constant term in the logistic regression equation, representing the value of the log-odds when all independent variables are zero.
- **Maximum Likelihood Estimation (MLE):** The method used to estimate the parameters of the logistic regression model by maximizing the likelihood function, which measures the probability of observing the data given the model parameters.
- **Confusion Matrix:** A table used to evaluate the performance of a classification model, showing the counts of true positives, true negatives, false positives, and false negatives.

b) Types of Logistic Regression: Logistic Regression can be classified into 3 categories. These are: -

- **Binomial:** There are only two conceivable forms of dependent variables in binomial logistic regression, such as 0 or 1, Pass or Fail, etc.
- **Multinomial:** The dependent variable in multinomial logistic regression may be one of three different unordered categories, such as "cat," "dogs," or "sheep".
- **Ordinal:** There are three or more possible ordered sorts of dependent variables in ordinal logistic regression, such as "low," "medium," or "high".

3.2.2 Random Forest Classification

Among the supervised learning techniques, Random Forest is a well-known machine learning algorithm. It can be used for machine learning problems involving both regression and classification. It is based on the idea of ensemble learning, a technique that combines multiple classifiers to solve a complex problem and improve the performance of the model.

Random Forest is a classifier that "contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."

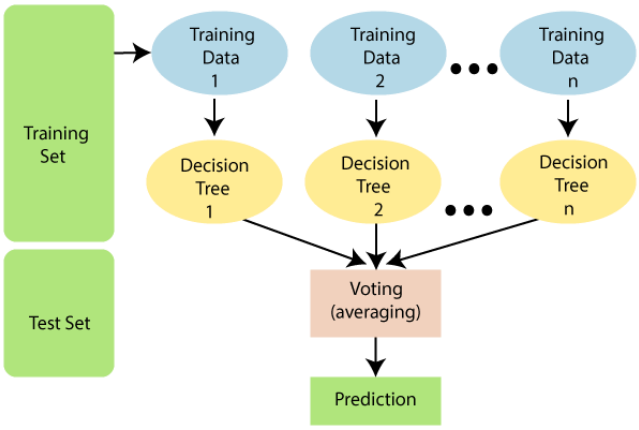

Figure 3.2: Random Forest Classification

The random forest gathers the forecasts from each decision tree instead than relying on a single one, and it predicts the outcome based on the majority vote of predictions. Higher accuracy and a decreased chance of overfitting are associated with larger tree counts in the forest. The Random Forest algorithm's operation is illustrated in figure 2.

3.2.3 Decision Tree

The decision tree is one of the best supervised learning strategies for both regression and classification issues. It creates a tree structure similar to a flowchart, with each internal node denoting a test on an attribute, each branch denoting the test's result, and each leaf node (terminal node) carrying a class name. The training data is divided into subsets based on attribute values iteratively until a halting condition such as the maximum tree depth or the minimum number of samples required to split a node is met.

a) **Terminologies involved in Decision Tree:** The following are a few terms that are frequently used in decision trees:

- **Root Node:** The root node in a decision tree is the highest node, representing the whole dataset and divided into child nodes according to feature values.
- **Internal Nodes:** Decision tree nodes that show splits based on feature values and from where child nodes branch.
- **Leaf Node (Terminal Node):** Nodes in a decision tree that indicate final decision outcomes or classifications and from which no child nodes branch off are called leaf nodes, sometimes known as terminal nodes.
- **Splitting Criterion:** The information gain, entropy, or Gini impurity criterion that is applied at each node to divide the data. It establishes the dividing feature and threshold.
- **Decision Rule:** Regulations at every internal node that specify how the data should be divided according to the values of attributes.
- **Feature:** Attributes or variables used for splitting the data at each node in the decision tree.
- **Branch:** The path from the root node to a leaf node, representing the sequence of decisions made based on feature values.
- **Pruning:** A technique used to prevent overfitting by removing unnecessary branches or nodes from the decision tree.
- **Entropy:** A measure of impurity or disorder in a dataset, used as a splitting criterion to maximize information gain.
- **Gini Impurity:** A measure of impurity or uncertainty in a dataset, used as a splitting criterion to minimize impurity in child nodes.
- **Information Gain:** A metric used to evaluate the effectiveness of a split in reducing uncertainty in classification tasks, calculated as the difference between the impurity of the parent node and the weighted impurity of the child nodes.
- **Decision Tree Depth:** The number of levels or layers in a decision tree, representing the complexity of the model and its ability to capture intricate patterns in the data.

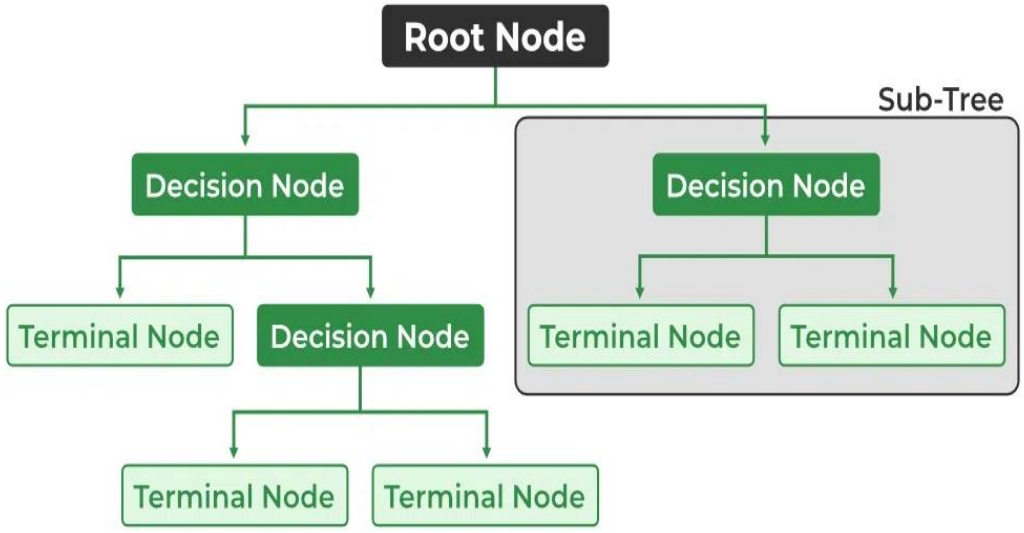

Figure 3.3: Decision Tree

3.2.4 Gradient Boosting

Gradient boosting is a robust boosting approach that turns several poor learners into strong learners by improving their performance. Gradient descent is used to train consecutive models to minimize a loss function, which can be anything from cross-entropy to mean squared error relative to the preceding model. In order to minimize the gradient of the loss function with respect to the predictions made by the current ensemble, this method generates a new weak model for each iteration. The new model's predictions are added to the ensemble and the process is repeated after it reaches a predetermined stopping condition.

Gradient Boosted Trees is a technique that uses CART (Classification and Regression Trees) as its base learner. The following diagram illustrates the gradient-boosted tree training procedure for regression problems.

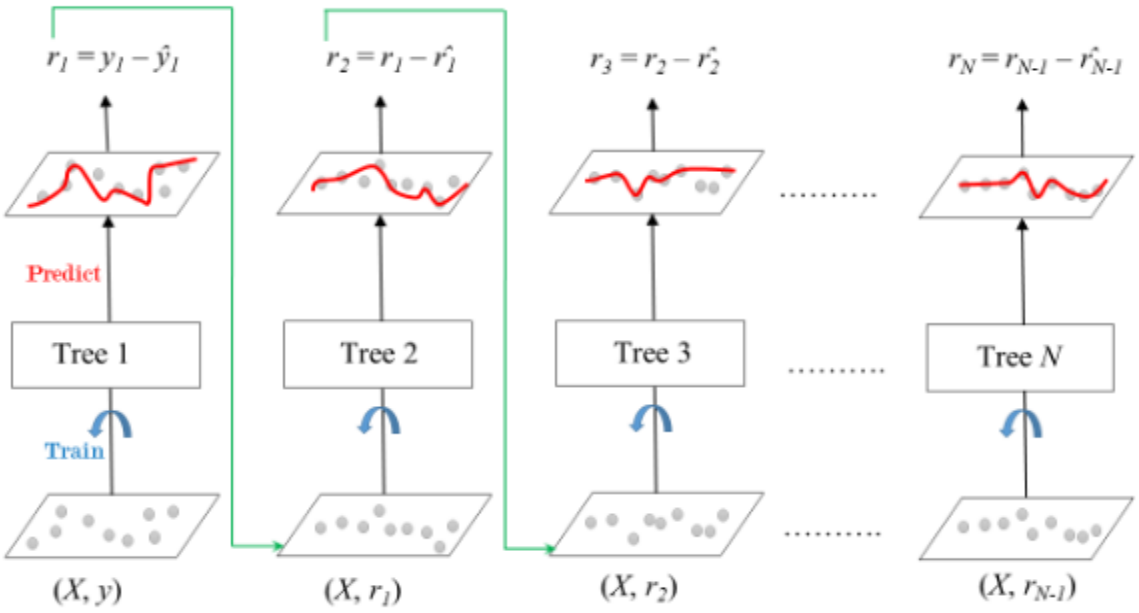

Figure 3.4: Gradient Boosted Trees for Regression

3.2.5 BERT (Bidirectional Encoder Representations from Transformers)

Using BERT for detecting and preventing deepfake news is a promising approach. BERT is a powerful natural language processing model that can analyse text and identify patterns. Here's how BERT can be applied in the context of deepfake news:

a) **Text Classification:** BERT can be trained to classify news articles or social media posts as real or potentially deepfake generated. By feeding it with labelled data containing both real and deepfake content, BERT can learn to differentiate between them based on language patterns and context.

b) **Fact-Checking:** BERT can assist in fact-checking by analysing the claims made in news articles or posts. It can cross-reference information with trusted sources and flag content that doesn't align with credible information.

c) **Context Analysis:** BERT can identify inconsistencies or anomalies in the language used within a piece of content. Deepfake news often contains subtle linguistic cues that may be missed by human readers but can be detected by models like BERT.

d) **Sentiment Analysis:** BERT can determine the sentiment of a news article or post. Some deepfake news might use emotionally charged language to manipulate readers, and sentiment analysis can help spot such manipulation.

e) **Monitoring social media:** BERT can be employed to continuously monitor social media platforms for the spread of deepfake news. It can identify trending topics and posts that are gaining traction, helping to address the issue promptly.

f) **Content Moderation:** BERT can assist in content moderation by flagging or removing deepfake content from websites and social media platforms.

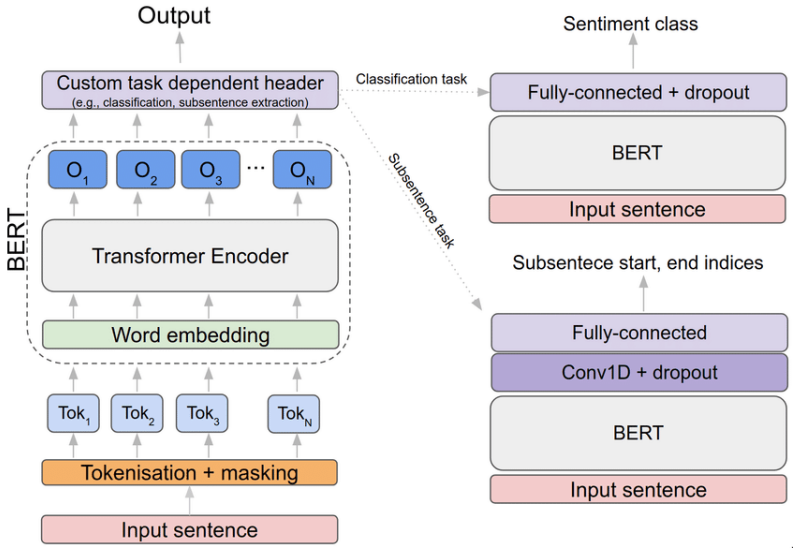

Figure 3.5: Bidirectional Encoder Representations from Transformers (BERT)

However, it's important to note that while BERT and similar models can be valuable tools in the fight against deepfake news, they are not foolproof. Deepfake technology is continually evolving, and creators are finding ways to make their content more convincing. Therefore, a combination of machine learning models like BERT, human fact-checkers, and efforts to promote media literacy are essential for an effective strategy to detect and prevent deepfake news.

CHAPTER 04

IMPLEMENTATION

4.1 Implementation

We've devised a research methodology comprising six steps to tackle the challenge of combating deepfake news in Bengali and identifying manipulated content within social media using machine learning.

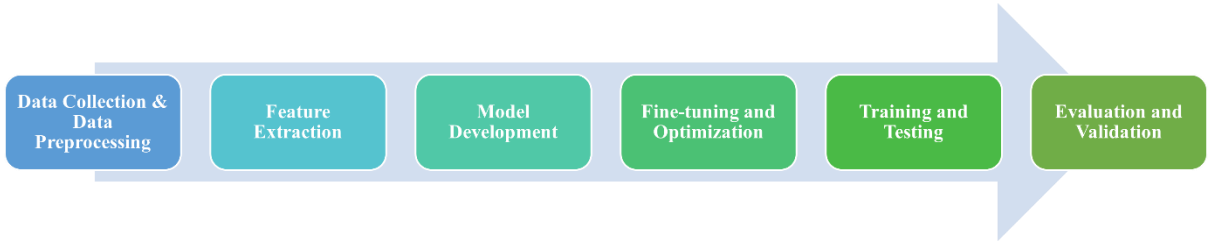

Figure 4. 1: Workflow Diagram to Combat Bangla Deepfake News

4.2 Data Collection & Dataset Description

A dataset refers to a collection of data files where tabular datasets may contain one or more database tables. In a table, each row signifies an occurrence, while each column denotes a distinct variable. Multiple files or documents collectively form a dataset.

A major challenge in identifying false news in Bengali was the dearth of pertinent research and resources in the Bengali language sector. Bengali did not have the same amount of datasets as better researched languages like English, Spanish, French, or Chinese, which is necessary to build reliable algorithms for detecting fake news. This lack of data turned into a significant barrier, similar to trying to reach an impossible goal. The lack of research on the identification of fake news in Bengali made it more difficult to obtain a suitable dataset for our study.To address this issue, we set out to find an appropriate dataset that was customized to the intricacies of the Bengali language. The process of obtaining a reliable dataset required careful work and thorough searches

via numerous news portals and websites. In spite of the rarity, we were able to locate an annotated dataset that had been painstakingly selected by a researcher from SUST, Bangladesh [32].

This collection of facts, which we assembled from several sources, served as the foundation for our research. Carefully selected from a wide range of news portals and websites, it provided a representative and varied sample of Bengali content, which was crucial for honing our false news identification algorithms. Every item in this dataset captured the core of the news scene in Bengali as well as the grammatical subtleties of the language. The extensive scope of our study not only served as a basis for our research, but it also shed light on possible solutions to the problem of scarce data in minority languages for upcoming research projects in this field. The filename and features of our collected dataset are depicted in the table 2.

Table 2: Contents of Dataset

Filename	**Authetic-48K.csv**	**LabelAuthentic7K.csv**	**Fake-1K.csv**	**LabelFake-1K.csv**
Rows	48K	7k	1.3K	1.3K
Columns	7	7	7	7
Features	ArticleID, Domain, Date, Category, Headline, Content, Label			

Within our research paper, we utilized two out of four Microsoft Excel .csv files, namely Authentic48K.csv and Fake-1K.csv. This dataset is labeled, comprising approximately 48,000 authentic news samples and roughly 1,300 fake news instances.

We collected another authentic dataset. This dataset contains 2K news and the title of the file is BanAuthFakeNews.csv. The features of this dataset are depicted in the table 3.

Table 3: Contents of Dataset

Filename	**BanAuthFakeNews.csv**
Rows	2K
Columns	5
Features	ArticleID, Category, Headline, Content, Label

The description of the column of the dataset are shown in below table 4.

Table 4: Description of Column

Column Name	Description
ArticleID	Serial Number of the News Article
Domain	Address of News Article
Date	News Published Date
Category	Category of the News Article
Headline	Headline of the News Article
Content	Description of the News Article
Label	Indicates whether the content is Authentic or Fake.

The sample of Authentic dataset are shown in the below table 5.

Table 5: Sample of Authentic Dataset

ArticleID	8
Domain	http://ittefaq.com.bd
Date	2018-09-20 16:50:33
Category	National
Headline	ডিজিটাল পাঠ্যবই শিক্ষার্থী ও শিক্ষক উভয়ের জন্য সহায়ক হবে: শিক্ষামন্ত্রী
Content	শিক্ষামন্ত্রী নুরুল ইসলাম নাহিদ বলেছেন, ইন্টারএকটিভ ডিজিটাল টেক্সটবুক (আইডিটি) শিশুদের পাঠ গ্রহণে সহায়ক হবে। তিনি বলেন, আইডিটি শিক্ষার্থীদের সহজে বুঝতে ও শিখতে যেমন সহায়তা করবে, তেমনি শিক্ষকদের জন্যও এটি সহায়ক হবে। আজ বৃহস্পতিবার রাজধানীর সেগুনবাগিচায় আন্তর্জাতিক মাতৃভাষা ইনস্টিটিউট মিলনায়তনে ৬ষ্ঠ শ্রেণির পাঠ্যবই ‘ইন্টারএকটিভ ডিজিটাল টেক্সটবুকের (আইডিটি)’ উদ্বোধন অনুষ্ঠানে প্রধান অতিথির বক্তৃতায় একথা বলেন শিক্ষামন্ত্রী। শিক্ষামন্ত্রী বলেন, ৬ষ্ঠ শ্রেণির ১৬টি ইন্টারএকটিভ পাঠ্যবই তৈরি করা হয়েছে। বইগুলোতে অডিও, ভিডিও, টেক্সট এবং এনিমেশন ব্যবহার করা হয়েছে। এতে বিষয়বস্তু শিক্ষার্থীদের কাছে আকর্ষনীয় ও আনন্দদায়ক হবে। সহজে এসব বই থেকে তারা কাঙ্খিত পাঠ গ্রহণ করতে পারবে।ফলে শিক্ষার্থীদের শিখন স্থায়ী হবে। নাহিদ বলেন, এর আগে নবম-দশম শ্রেণির ৬টি বইয়ের ই-লানিং ম্যাটেরিয়াল তৈরি করা হয়েছে। সেসিপ প্রকল্পের আওতায় ৭ম শ্রেণির ৬টি এবং ৮ম শ্রেণির ৬টি ইন্টারএকটিভ ডিজিটাল বই তৈরি প্রক্রিয়াধীন রয়েছে। শিক্ষামন্ত্রী বলেন, বিশ্বমানের শিক্ষা অর্জনে আধুনিক জ্ঞান ও প্রযুক্তিকে ধারণ করতে হবে। প্রযুক্তির সুযোগগুলো কাজে লাগাতে হবে। তিনি বলেন, শিক্ষায় তথ্যপ্রযুক্তির ব্যবহারে বিভিন্ন পদক্ষেপ নেয়া হয়েছে। ষষ্ঠ শ্রেণি থেকে আইসিটি শিক্ষা বাধ্যতামূলক করা হয়েছে। শিক্ষাক্ষেত্রে প্রযুক্তির ব্যবহার অনেক বৃদ্ধি পেয়েছে। ভর্তি ও ফলাফল প্রকাশসহ শিক্ষা মন্ত্রণালয়ের অনেক কার্যক্রম পেপারলেস হয়েছে। জাতীয় শিক্ষাক্রম ও পাঠ্যপুস্তক বোর্ডের চেয়ারম্যান প্রফেসর নারায়ণ চন্দ্র সাহার সভাপতিত্বে এ অনুষ্ঠানে মাধ্যমিক ও উচ্চ শিক্ষা বিভাগের সচিব মো. সোহরাব হোসাইন, কারিগরি ও মাদ্রাসা শিক্ষা বিভাগের সচিব মো. আলমগীর, এশীয় ডেভেলপমেন্ট ব্যাংকের (এডিবি) কান্ট্রি ডিরেক্টর মনমোহন পারকাশ এবং টিচিং কোয়ালিটি ইমপ্রুভমেন্ট-২ প্রকল্পের পরিচালক মো. জহির উদ্দিন বাবর বক্তৃতা করেন।
Label	1

The sample of Fake dataset are shown in the below table 6.

Table 6: Sample of Fake Dataset

ArticleID	4
Domain	channeldhaka.news
Date	2018-06-30T15:56:47+00:00
Category	Sports
Headline	অবসর নেয়ার ঘোষণা দিলেন মেসি !
Content	রাশিয়া বিশ্বকাপ নকআউট পর্বে ফ্রান্সের সাথে ৪-৩ গোলে পরাজয়ের গ্লানিতে ফুটবল থেকে অবসরের ঘোষণা দিলেন মেসি। খেলা শেষে মাঠ ছাড়তে ছাড়তে বিড়বিড় করে এমনটাই বলছিলেন আর্জেন্টাইন এই সুপারস্টার !ভালোই চলছিলো ম্যাচ। শুরুতে ফ্রান্স গোল দিলে মাঠ কাঁপালেও পরবর্তীতে দুটি গোল যুক্ত হয় আর্জেন্টিনার ঝুলিতে। বুক চেতিয়ে খেলা শুরু করেন মেসি। হঠাৎই বদলে যায় ম্যাচের দৃশ্যপট। একে একে ফ্রান্স আরো তিনটি গোল ঢুকিয়ে দেয় আর্জেন্টিনার জালে। এরই সাথে বিশ্বকাপ থেকে বিদায় ঘন্টা বেজে যায় আর্জেন্টিনার!মাঠ ছাড়ার সময় গেটের চিপায় থাকা এক সাংবাদিক দেখতে পান, ক্লান্তি অবসাদপূর্ণ শরীরে মাঠ ছাড়তে ছাড়তে মেসি মাথা নিচু করে কি যেনো বলছেন। ওই সাংবাদিক মেসির লিপ রিডিং করে বুঝতে পারেন, মেসি বলছে, “আমি আর ফুটবল খেলবোনা। আমি আর ফুটবল খেলবোনা।”বাতাসের বেগে কথাটি স্টেডিয়ামজুড়ে ছড়িয়ে পরে। পরবর্তীতে মেসির অবসর নেয়ার ব্যাপারে দ্রুত এ খবরটি স্টেডিয়াম প্রাঙ্গনে ভাইরাল হয়ে যায়। এবং অনেকেই তা বিশ্বাস করে বসেন। যদিও আনুষ্ঠানিকভাবে মেসি এখনো কিছু জানাননি !এখন শুধুই এ কিংবদন্তীর মুখ থেকে অবসর নেয়ার ঘোষণাটি শোনার পালা…
Label	0

Although this dataset is labeled, further processing is needed before it can be used to train the model to identify bogus news in Bengali.

4.2.1. Data Preprocessing

Data preprocessing involves a series of steps and techniques used to clean, transform, tokenize and organize raw data before it's used in analysis or model training. It includes tasks like handling missing values, normalization, scaling, encoding categorical variables, and other operations to make the data more suitable for machine learning algorithms or analytical processes. The data was tokenized using two different techniques. We trained the BERT model to tokenize sentences using WordPiece tokenization, which is very useful for managing new vocabulary. HuggingFace's publicly available Transformers implementation was used to setup our dataset for tokenization. Next, we made use of a pre-trained transformer model that was highly efficient and based on a tokenization library created especially for the Bengali language.

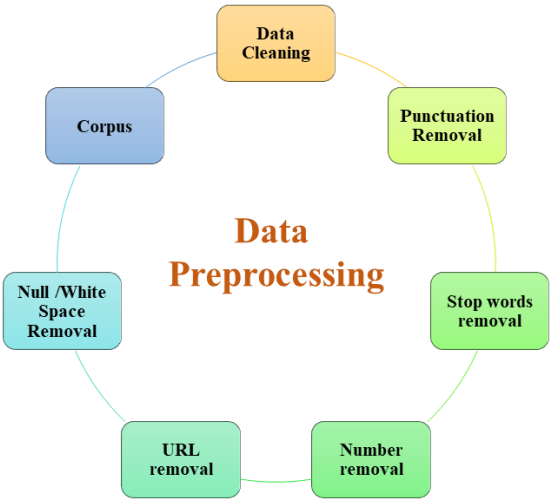

Figure 4. 2: Data Preprocessing

Data from the real world isn't always accurate and constant. These data are inaccurate and don't show particular patterns or behaviors. Preprocessing data has the ability to fix these problems. In the realm of Machine Learning, data preprocessing modifies the dataset to a format that algorithms can easily understand and handle during the procedures. The data preprocessing involved several key steps:

- **Data Cleaning:** Removed punctuation, numbers, null values, and duplicates from the raw dataset.
- **Stop Words Removal:** Filtered out words with minimal meaning.
- **Formatting:** Transformed the data into two distinct formats for clean and standardized analysis. The two formats are-
    i. **Corpus:** A corpus is a structured collection of texts used for linguistic analysis or research purposes. To convert the dataset into a corpus structure, we employed Pandas, a Python library utilized for data analysis.
    ii. **URL Removal***:* URLs are either reference links or metadata (such as HTML tags) if they are found in the content, but they don't provide any useful information. First, we pre-processed the dataset by removing all URLs.

With all these steps wrapped up, our polished corpus is ready.

4.3. Feature Extraction

The process of converting unprocessed data into a format that machine learning algorithms can use efficiently is known as feature extraction. It involves selecting, combining, or transforming variables within the dataset to create new features that are more informative and representative of the underlying data patterns. By removing unnecessary details or characteristics from the data, dimensionality is decreased and the algorithm's capacity to learn and generate precise predictions is improved. In our corpus, we've employed **Term Frequency-Inverse Document Frequency (TF-IDF),** recognized as a top-notch method for extracting features. Term Frequency-Inverse Document Frequency assesses the significance of terms inside a document or corpus in order to extract features. Two factors are computed:

- **Term Frequency (TF):** Calculates a term's frequency of occurrence in a document. By calculating a word's frequency of occurrence in relation to the total number of words in a phrase, one can apply the Term Frequency (TF) technique.
- **Inverse Document Frequency (IDF):** Evaluates a term's uniqueness across all documents in the corpus. Rare words that appear in fewer documents tend to have higher IDF scores and are considered more significant.

The TF-IDF value results from multiplying TF by IDF. A higher TF-IDF signifies the rarity of the word or phrase.

4.4. Model Development

Our aim was to select the most efficient approach in our model for detecting fake news. We can see the proposed model in Figure 8.

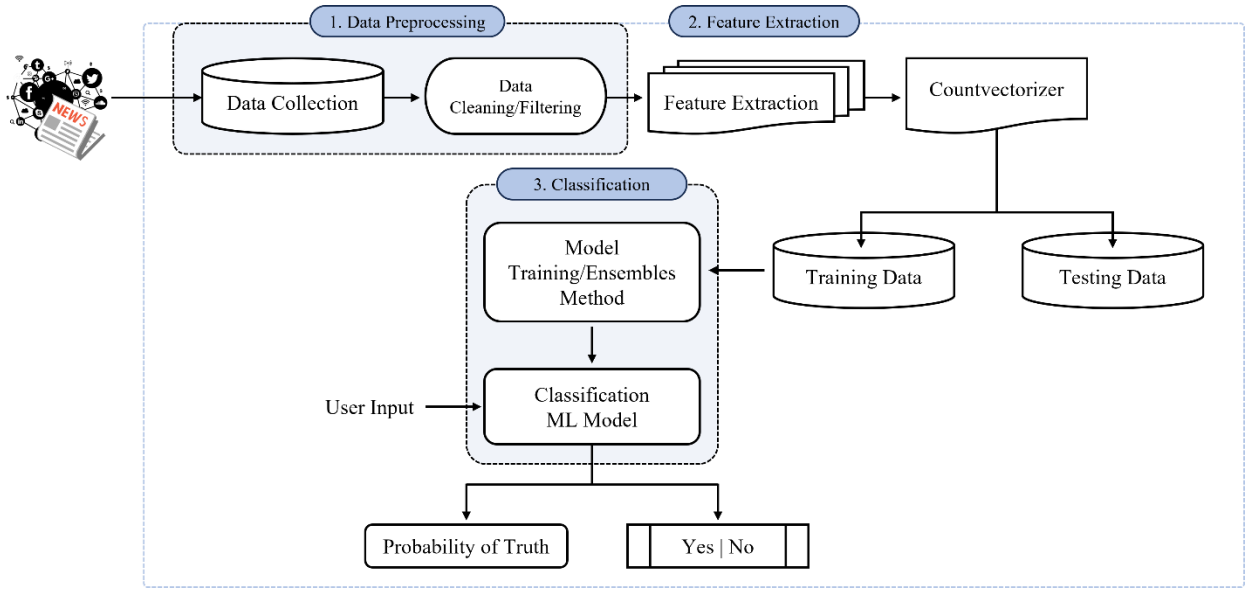

Figure 4. 3: Proposed Method for Bangla Fake News Detection

After collecting the initial dataset, the subsequent step involved data processing essential for feeding machine learning classifiers and extracting features. This process is crucial for various text classification tasks as it enables the extraction of relevant information from the dataset, facilitating accurate analysis and classification of textual data. Essentially, it ensures that the data is structured and transformed in a way that allows machine learning algorithms to effectively interpret and derive meaningful insights.

The preprocessing phase demanded extensive tasks like eliminating punctuation marks, stop words, URLs, numbers, converting cases, identifying named entities, stemming, and lemmatizing. However, given the constraints of working with the Bengali language and limited tools, we focused on removing punctuation marks, stop words, and numbers to cleanse the dataset. We utilized TF-IDF for feature extraction.

After completing the initial steps, we split our dataset into two portions: 70% for training the machine learning classifier and 30% for testing its performance. This division allows us to train the model on a substantial portion of the data and evaluate its effectiveness on unseen data.

We experimented with many classifiers, including Decision Tree (DT), Gradient Boosting (GB), Random Forest (RF), Logistic Regression (LR), and Bidirectional Encoder Representations from Transformers (BERT). Each of these classifiers has unique characteristics and ways of learning from the data. By testing these different models, we aimed to identify which one performs best for our task of detecting fake news in Bengali.

After training and testing each classifier, the model evaluated their performances. The goal was to determine which classifier yielded the most accurate and reliable results in distinguishing between real and fake news in Bengali. The selected classifier was the one that demonstrated the highest accuracy and robustness in identifying fake news within the dataset. This approach enabled us to finalize the best-performing classifier to use for detecting fake news in Bengali language content.

4.5 Fine-tuning and Optimization

Fine-tune the selected models to improve their performance on Bengali-specific deepfake detection. Explore techniques such as

- Transfer learning
- Domain adaptation or
- Model assembling

to enhance the models' robustness and adaptability to Bengali language nuances and cultural contexts.

4.6 Evaluation and Validation

Evaluate the performance of the developed models using appropriate evaluation metrics such as

i. Accuracy
ii. Recall

iii. Precision and

iv. F1 score.

A. **Accuracy:** To Calculate the accuracy of the performance model the following equation can be used.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad \text{..... 4.1}$$

F1 -Score represent the trade off between precision and recall. It calculates the harmonic mean between each of the two. Thus, it takes both the false positive and the false Negative observations into account. F1-Score can be calculated using the following formula.

Validate the models using cross-validation techniques or separate validation datasets. Compare the performance of different models and identify the most effective approaches for deepfake detection in Bengali.

B. **Recall:** In machine learning, recall is a performance metric that measures the ability of a model to correctly identify all relevant instances of a particular class or category within a dataset. It is a measure of the model's ability to avoid false negatives. Recall is particularly important in scenarios where the identification of certain instances is critical, such as in medical diagnosis or fraud detection.

Mathematically, recall is calculated as the ratio of the number of true positives (TP) to the sum of true positives and false negatives (FN):

$$\text{Recall} = TP / (TP + FN) \quad \text{..... 4.2}$$

True positives (TP) represent the instances that are correctly identified as positive by the model, while false negatives (FN) represent the instances that are incorrectly classified as negative but actually belong to the positive class.

A high recall value indicates that the model is effective at capturing a large proportion of the positive instances within the dataset. Conversely, a low recall value suggests that the model may be missing a significant number of positive instances.

While recall is an important metric, it should be considered alongside other performance measures, such as precision, accuracy, and F1 score, to gain a comprehensive understanding of a model's performance.

C. **Precision:** In machine learning, precision is a performance metric that measures the accuracy of the positive predictions made by a model. It quantifies the proportion of correctly predicted positive instances out of all instances predicted as positive by the model. Precision is particularly useful in scenarios where the focus is on minimizing false positives, such as in spam detection or anomaly detection.

   Mathematically, precision is calculated as the ratio of true positives (TP) to the sum of true positives and false positives (FP):

$$\text{Precision} = TP / (TP + FP) \quad \text{……..…… 4.3}$$

   True positives (TP) represent the instances that are correctly identified as positive by the model, while false positives (FP) represent the instances that are incorrectly classified as positive but actually belong to the negative class.

   A high precision value indicates that the model has a low rate of falsely classifying negative instances as positive. It suggests that when the model predicts a positive instance, it is highly likely to be correct. On the other hand, a low precision value indicates that the model has a higher rate of false positives, meaning that it is incorrectly labeling negative instances as positive.

   Precision should be considered alongside other performance measures, such as recall, accuracy, and F1 score, to gain a comprehensive understanding of a model's performance.

Depending on the specific problem and requirements, there may be a trade-off between precision and recall.

D. Performance Metrics:

To evaluate the performance of algorithms, I use different metrics. Most of them are based on the confusion matrix. A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance. The matrix displays the number of

i. True Positives (TP)
ii. True Negatives (TN)
iii. False Positives (FP)
iv. False Negatives (FN)

produced by the model on the test data.

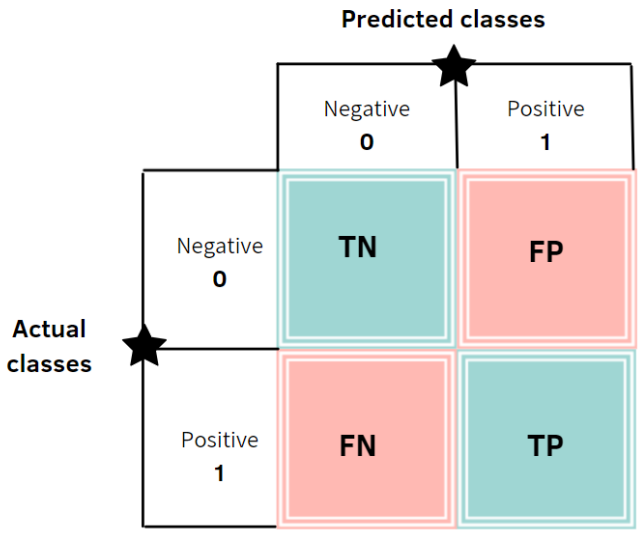

Figure 4. 4: Confusion Matrix

CHAPTER 05

RESULTS AND DISCUSSION

5.1 Result Analysis and Discussion

We will evaluate our model using a variety of measures, including accuracy, precision, recall, and F-1 score for each classifier. We divided our dataset into two parts: 25% for testing and 75% for training. In the upcoming sections, we will provide a thorough analysis of the performance exhibited by each classifier during testing.

5.2 Logistic Regression

The accuracy, precision, recall and F1-score of Logistic Regression algorithm are shown in below table 7.

Table 7: Accuracy, Precision, Recall and F1-score of Logistic Regression

Accuracy	97.71%
Precision	97.58%
Recall	97.71%
F1-score	96.98%

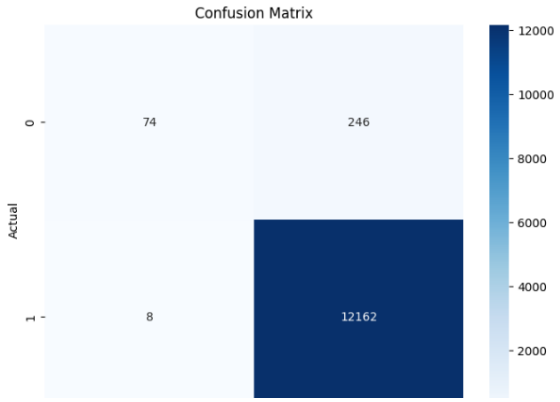

Figure 5. 1: Confusion Matrix for Logistic Regression

Figure 9 shows us that accuracy is about 97.71%. A score of 97.58% was awarded for precision. We discovered that 121162 authentic news items were correctly classified as authentic news, and 74 false news items were successfully classified as fake news, demonstrating the accuracy of our forecasts. Moreover, the classifier's validity is demonstrated by its 97.71% recall and 96.98% F-1 score.

5.3 Decision Tree

The accuracy, precision, recall and F1-score of Decision Tree algorithm are shown in table 8.

Table 8:Accuracy, Precision, Recall and F1-score of Decision Tree

Accuracy	97.62%
Precision	97.42%
Recall	97.62%
F1-score	97.51%

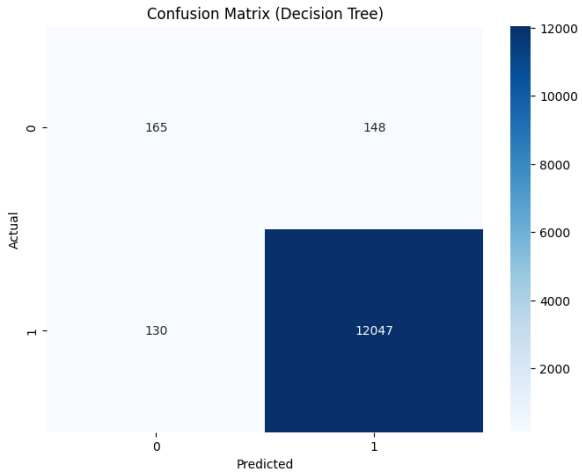

Figure 5. 2: Confusion Matrix for Decision Tree

Figure 10 shows us that accuracy is about 97.62%. A score of 97.42% was awarded for precision. We discovered that 12047 authentic news items were correctly classified as authentic news, and 165 false news items were successfully classified as fake news, demonstrating the accuracy of our forecasts. Moreover, the classifier's validity is demonstrated by its 97.62% recall and 97.51% F-1 score.

5.4 Gradient Boosting

The accuracy, precision, recall and F1-score of Gradient Boosting algorithm are shown in table 9.

Table 9: Accuracy, Precision, Recall and F1-score of Gradient Boosting

Accuracy	98.11%
Precision	97.93%
Recall	98.11%
F1-score	97.75%

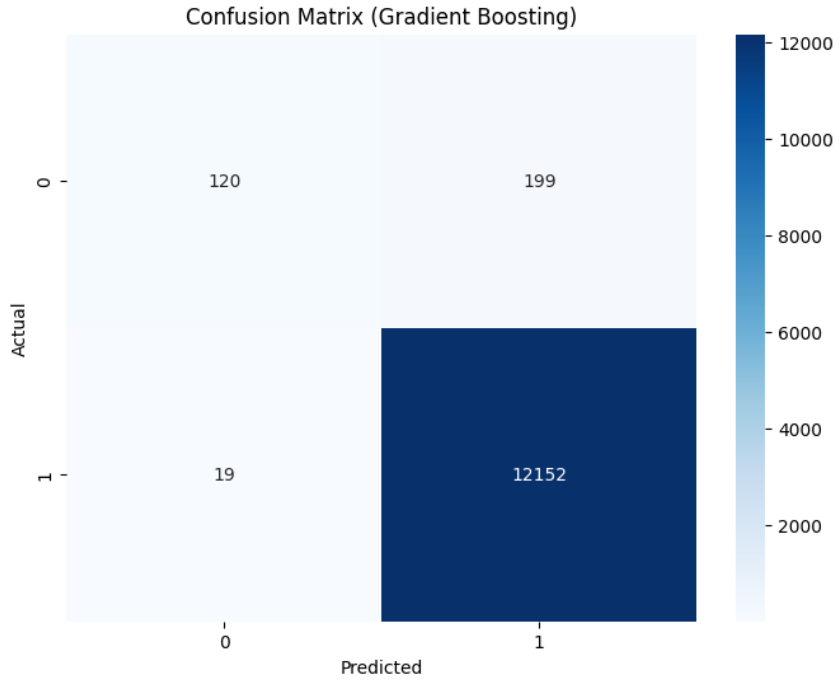

Figure 5. 3: Confusion Matrix for Gradient Boosting

Figure 11 shows us that accuracy is about 98.11%. A score of 97.93% was awarded for precision. We discovered that 12152 authentic news items were correctly classified as authentic news, and 120 false news items were successfully classified as fake news, demonstrating the accuracy of our forecasts. Moreover, the classifier's validity is demonstrated by its 98.11% recall and 97.75% F-1 score.

5.5 Random Forest

The accuracy, precision, recall and F1-score of Decision Tree algorithm are shown in table 10.

Table 10: Accuracy, Precision, Recall and F1-score of Random Forest

Accuracy	97.85%
Precision	97.90%
Recall	97.85%
F1-score	97.20%

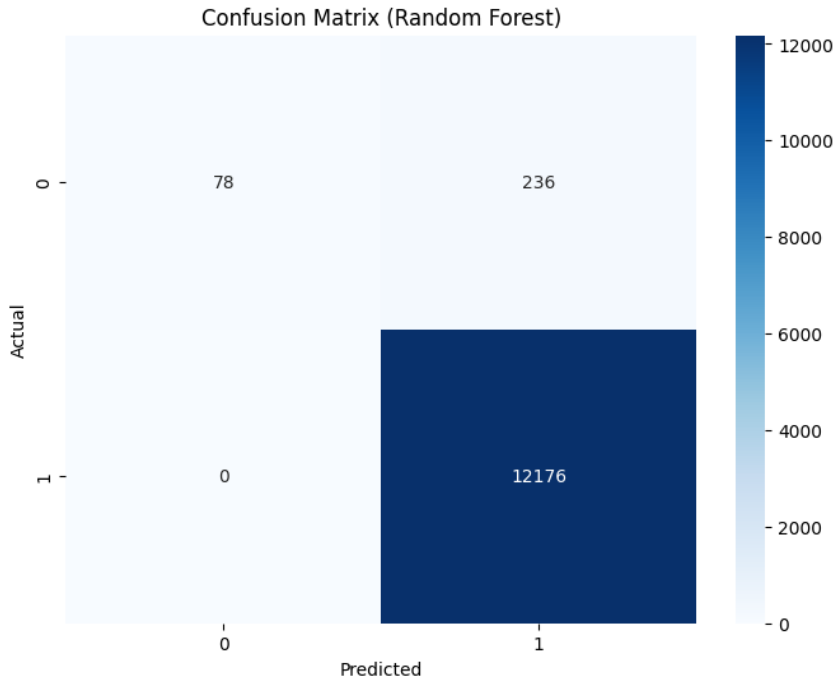

Figure 5. 4: Confusion matrix for Random Forest

Figure 12 shows us that accuracy is about 97.85%. A score of 97.90% was awarded for precision. We discovered that 12176 authentic news items were correctly classified as authentic news, and 78 false news items were successfully classified as fake news, demonstrating the accuracy of our forecasts. Moreover, the classifier's validity is demonstrated by its 97.85% recall and 97.20% F-1 score.

5.6 BERT (Bidirectional Encoder Representations from Transformers)

The accuracy, precision, recall and F1-score of BERT algorithm are shown in table 11.

Table 11:Accuracy, Precision, Recall and F1-score of BERT

Accuracy	98.52%
Precision	98.56%
Recall	99.95%
F1-score	99.25%

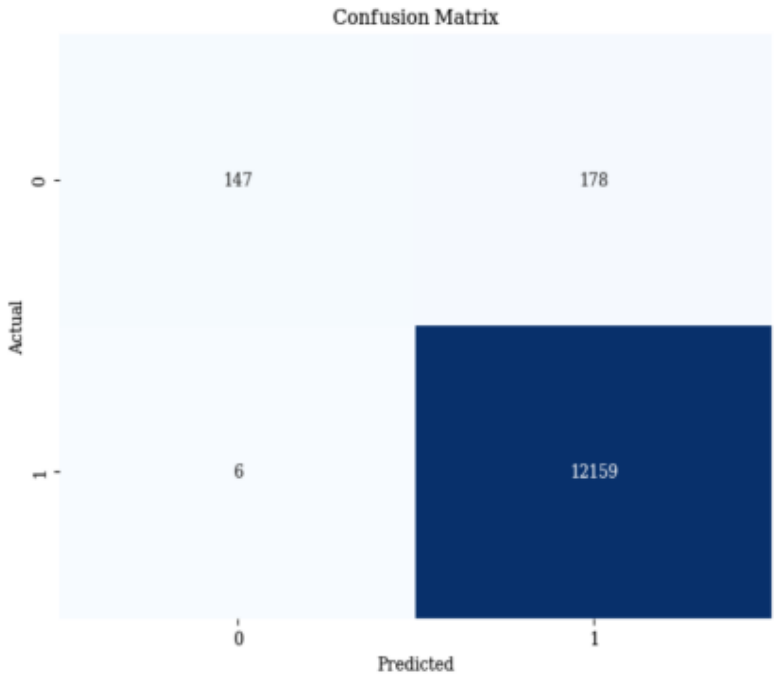

Figure 5. 5: Confusion Matrix for BERT

Figure 13 shows us that accuracy is about 98.52%. A score of 98.56% was awarded for precision. We discovered that 12159 authentic news items were correctly classified an authentic news, and 147 false news items were successfully classified as fake news, demonstrating the accuracy of our forecasts. Moreover, the classifier's validity is demonstrated by its 99.95% recall and 99.25% F-1 score.

5.6 Comparison

The summery of these classifier is shown the below table 12.

Table 12: Comparison summary of the classifiers

Model Name	Accuracy	Precision	Recall	F1-score
Logistic Regression	97.11%	97.58%	97.71%	96.98%
Decision Tree	97.62%	97.42%	97.62%	97.51%
Gradient Boosting	98.11%	97.93%	98.11%	97.75%
Random Forest	97.85%	97.90%	97.85%	97.20%
BERT	**98.52%**	**98.56%**	**99.95%**	**99.25%**

Table 12 indicates that the Gradient Boosting Classifier achieved an accuracy of 98.11%. Random Forest and Decision Tree Classifier also showed outstanding accuracy of 97.85% and 97.62% respectively. However, with 98.52% accuracy, the BERT Classifier had the highest performance. We may also observe that, due to the algorithm's design, the Logistic Regression Classifier had the lowest accuracy of 97.11%. The column chart in Figure 14 illustrates the results we obtained using the classifiers.

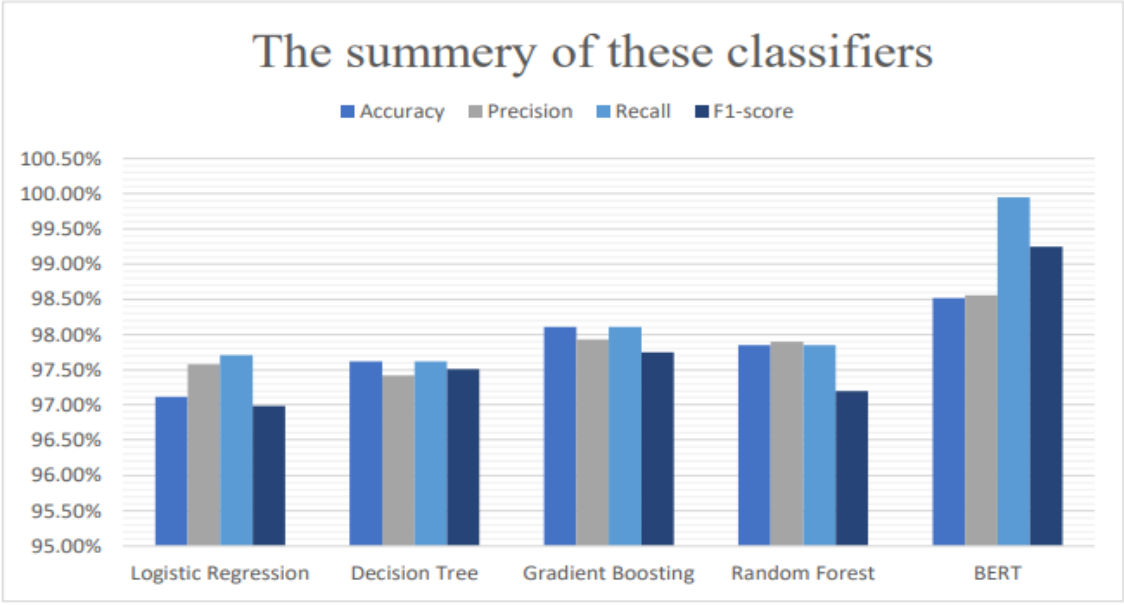

Figure 5. 6: Histogram of the Classifiers

CHAPTER 06

CONCLUSION

6.1 Conclusion

Proficient machine learning techniques for detecting fake news are commonly proposed in the English language. However, there has been limited scholarly research conducted in the Bengali language to address this issue. Furthermore, constructing a substantial corpus of annotated datasets for Bengali news poses another significant challenge. Nevertheless, this study utilized a recently published annotated dataset in Bengali, meticulously labeled by professionals. The analysis reveals that all classifiers demonstrated strong performance in distinguishing between authentic and fake news, with BERT exhibiting the highest overall performance. Gradient Boosting also exhibited notable accuracy along with balanced precision and recall. The selection of the optimal classifier may vary depending on specific requirements and considerations. Overall, the research underscores the effectiveness of these algorithms in news classification, offering valuable insights for combating misinformation and upholding information integrity.

Future Work

REFERENCES

1) An importance of an online news portal, http://www.syvjournal.com/an - importance-of-an-online-news-portal, 2021.

2) S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science,* vol. 359, no. 6380, pp. 1146–1151, 2018. doi: 10.1126/science.aap9559. eprint: https://www.science.org/doi/pdf/10.1126/science.aap9559.

3) I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity,* vol. 2020, pp. 1–11, Oct. 2020. doi: 10.1155/2020/8885861.

4) E. M. Mahir, S. Akhter, M. R. Huq, et al., "Detecting fake news using machine learning and deep learning algorithms," in 2019 *7th International Conference on Smart Computing & Communications (ICSCC), IEEE,* 2019, pp. 1–5.

5) Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).*

6) Manik, J. A. (2012). A hazy picture appears. https://www.thedailystar.net/newsdetail-252212, October. [Online; posted 03- October-2012].

7) Long, Y., Lu, Q., Xiang, R., Li, M., and Huang, C.-R. (2017). Fake news detection through multi-perspective speaker profiles. *In Proceedings of the Eighth International Joint Conference on Natural Language Processing* (Volume 2: Short Papers), pages 252–256, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

8) Yang, F., Mukherjee, A., and Dragut, E. (2017). Satirical news detection and analysis using attention mechanism and linguistic features. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* pages 1979–1989, Copenhagen, Denmark, September. Association for Computational Linguistics.

9) Karadzhov, G., Nakov, P., Marquez, L., Barr ` on-Cede ´ no, A., ˜ and Koychev, I. (2017). Fully automated fact checking using external sources. *In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017,* pages 344–353, Varna, Bulgaria, September. INCOMA Ltd.

10) Dong, X., Victor, U., Chowdhury, S., and Qian, L. (2019). Deep two-path semi-supervised learning for fake news detection. *CoRR*, abs/1906.05659.

11) Alawadh, H. M., Alabrah, A., Meraj, T., & Rauf, H. T. (2023). Attention-Enriched Mini-BERT Fake News Analyzer Using the Arabic Language. *Future Internet,* 15(2), 44

12) Soll, J.; White, J.B.; Sitrin, S.S.; Carly.; Gerstein, B.M. The Long and Brutal History of Fake News. *Politico Magazine,* 18 December 2016

13) Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsl.* 2017, 19, 22–36.

14) Schonfeld, E. Citizen “Journalist” Hits Apple Stock with False (Steve Jobs) Heart Attack Rumor. 2008. https://techcrunch.com/2008 /10/03/citizen-journalist-hits-apple-stock-with-false-steve-jobs-heart-attack-rumor (accessed on 15 May 2022).
15) Zhou, X.; Zafarani, R. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explor. Newsl.* 2019, 21, 48–60.
16) L. Alsudias and P. Rayson, “Covid-19 and arabic twitter: How can arab world governments and public health organizations learn from social media?” in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL* 2020, 2020
17) Muaad, A.Y.; Jayappa Davanagere, H.; Benifa, J.; Alabrah, A.; Naji Saif, M.A.; Pushpa, D.; Al-Antari, M.A.; Alfakih, T.M. Artificial intelligence-based approach for misogyny and sarcasm detection from Arabic texts. *Comput. Intell. Neurosci.* 2022, 2022
18) Mughaid, A.; Al-Zu’bi, S.; Al Arjan, A.; Al-Amrat, R.; Alajmi, R.; Zitar, R.A.; Abualigah, L. An intelligent cybersecurity system for detecting fake news in social media websites. *Soft Comput.* 2022, 26, 5577–5591
19) Gumaei, A.; Al-Rakhami, M.S.; Hassan, M.M.; De Albuquerque, V.H.C.; Camacho, D. An effective approach for rumor detection of Arabic tweets using extreme gradient boosting method. *Trans. Asian Low-Resour. Lang. Inf. Process.* 2022, 21, 1–16.
20) Perez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, ´ R. (2018). Automatic detection of fake news. *In Proceedings of the 27th International Conference on Computational Linguistics,* pages 3391–3401, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.
21) Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter,* 19(1):22–36.
22) Wang, W. Y. (2017). ” liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:*1705.00648.
23) Amer, E.; Kwak, K.S.; El-Sappagh, S. Context-Based Fake News Detection Model Relying on Deep Learning Models. *Electronics* 2022, 11, 1255.
24) Lasotte, Y.; Garba, E.; Malgwi, Y.; Buhari, M. An Ensemble Machine Learning Approach for Fake News Detection and Classification Using a Soft Voting Classifier. *Eur. J. Electr. Eng. Comput. Sci.* 2022, 6, 1–7.

25) N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," *in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management,* 2017, pp. 797–806.

26) A. Benamira, B. Devillers, E. Lesot, A. K. Ray, M. Saadi, and F. D. Malliaros, "Semi-supervised learning and graph neural networks for fake s detection," *in 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE,* 2019, pp. 568–569.

27) Ahmed, H., Traore, I., & Saad, S. (2018). Detecting opinion spams and fake news using text classification. Security and Privacy, 1(1), e9.

28) Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic detection of fake news. arXiv preprint arXiv:1708.07104.

29) Monteiro RA, Santos RL, Pardo TA, De Almeida TA, Ruiz EE, Vale OA (2018) Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In: international conference on computational processing of the portuguese language, Springer, pp 324–334

30) Hanselowski, A., PVS, A., Schiller, B., Caspelherr, F., Chaudhuri, D., Meyer, C. M., & Gurevych, I. (2018). A retrospective analysis of the fake news challenge stance detection task. arXiv preprint arXiv:1806.05180.

31) Gaydhani A, Doma V, Kendre S, Bhagwat L.Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach.2018.

32) M. Z. Hossain, M. A. Rahman, M. S. Islam, and S. Kar, "Banfakes: A dataset for detecting fake s in bangla," *arXiv preprint arXiv:*2004.08789, 2020.