

Identifying Bangla Deceptive News Using Machine Learning and Deep Learning Algorithms

by

Anindya Roy Piyal

19101577

Shams Iqbal

19101578

Anupom Ray Rohan

19101483

Nowshin Zaman

19101018

Nowshin Meheja

19101035

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2023

© 2023. Brac University
All rights reserved.

Declaration

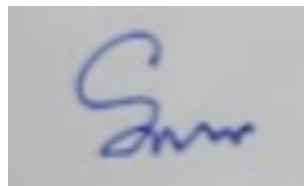
It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

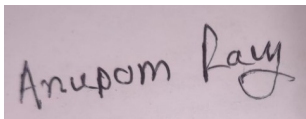
Student's Full Name & Signature:



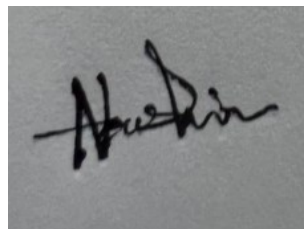
Anindya Roy Piyal
19101577



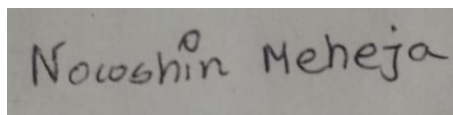
Shams Iqbal
19101578



Anupom Ray Rohan
19101483



Nowshin Zaman
19101018



Nowshin Meheja
19101035

Approval

The thesis titled “Identifying Bangla Deceptive News Using Machine Learning and Deep Learning models” submitted by

1. Anindya Roy Piyal(19101577)
2. Shams Iqbal(19101578)
3. Anupom Ray Rohan(19101483)
4. Nowshin Zaman(19101018)
5. Nowshin Meheja(19101035)

Of Spring, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 29, 2023.

Examining Committee:

Supervisor:
(Member)



Dr. Md. Khalilur Rhaman, PhD
Professor
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Robiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Internet-based resources are utilized by the vast majority of individuals today. The news published on websites and shared on social media platforms are examples of such resources. Due to the increasing number of content creators, online media portals, and news portals, it has become nearly impossible to verify the veracity of news headlines and undertake thorough assessments of them. The overwhelming majority of fraudulent headlines contain misleading or false information. They obtain more views and shares from people of all ages by using clickbait titles that contain fictitious terms or false information. However, these false and misleading headlines cause chaos in the lives of the average individual and mislead them in numerous ways. We have used recent Bangla news articles to create a model that can accurately determine the reliability of the news. In order to detect fake Bangla news stories, we have used approximately 10,000 news articles to train our machine learning and deep learning model. In addition, the Bengali language uses BNLP and BLTK for a wide range of natural language processing activities and `bn_w2v_wiki` a word embedding model for Bangla Language to represent words as vectors. The Synthetic Minority Oversampling Strategy (SMOTE) was used to remove the imbalance of our dataset. On the training data of our dataset, we have employed machine learning in addition to deep learning algorithm. Our deep learning model LSTM performs best with the accuracy of 91% . Also our machine learning model Random Forest and Support Vector Machine performs well enough to compete with LSTM for the prediction of fake news. The other machine learning algorithms included are LR, KNN, GNB, bagging, boosting. Furthermore, we have developed a website that takes Bangla news text as input and classifies the news with the help of our trained model. We believe our study will go a long way towards establishing a foundation in the research field of low resourced Bangla Language and open new door to future study.

Keywords: Fake-news, Bangla fake-news, BNLP, BLTK, `bn_w2v_wiki`, SMOTE, Machine Learning, LSTM, RFC, Deep Learning.

Acknowledgement

We are first and foremost grateful to Allah for his benevolence. Additionally, our parents who helped us complete our task.

Then, we wish to express our sincere appreciation to Sayantan Roy Arko sir for his invaluable contribution to our research. His insightful contributions and assistance with the implementation of machine learning and deep learning algorithms were crucial to the success of our research. We are extremely appreciative of his unwavering support and dedication throughout the research process.

In addition, we appreciate Mohiuddin Iqbal's assistance in constructing an aesthetically pleasing diagram for our thesis. His graphic design abilities and attention to detail made the visual representation of our research findings more accessible and engaging for readers. We greatly value his contribution and the effort he exerted to create a visually compelling representation of our work.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
List of Figures	1
1 Introduction	2
1.1 Inspiration for the Forecasting Model	2
1.2 Research Problem	3
2 Related Work	6
3 Methodolgy and Requirement Analysis	9
3.1 Data Collection	9
3.2 Data Preprocessing	10
3.2.1 Remove Punctuations	11
3.2.2 Remove Stopwords	11
3.2.3 Stemming	12
3.3 Feature Extraction	12
3.3.1 Word2vec	13
3.3.2 Bangla FastText Model	13
3.3.3 Gensim	14
3.4 Smote	15
4 A comprehensive look at the dataset	16
5 AI Models and Implementations	21
5.1 LSTM	22
5.2 Logistic Regression	23
5.3 Random Forest	24
5.4 KNN	25
5.5 Support Vector Machine	27
5.6 Gaussian Naive Bayes	28
5.7 Bagging	28

5.8	Boosting	29
6	Result Analysis	30
6.1	Accuracy, precision, recall, and F1:	30
6.2	Confusion Matrix:	33
6.3	ROC Curve:	34
6.4	Learning Curve:	38
7	Conclusion	41
7.1	Challenges	41
7.2	Limitations	42
7.3	Discussion and Future Work	43
	Bibliography	48

List of Figures

1.1	Roadmap	3
3.1	Architecture of our system	10
3.2	Preprocessing stages	10
3.3	Feature extraction methods	13
3.4	Glimpse of a 300d vector	15
3.5	Smote Technique	15
4.1	Authentic news sources count.	17
4.2	Authentic news categories	18
4.3	Fake news categories	18
4.4	Authentic news word cloud	19
4.5	Fake news word cloud	19
4.6	Data Preprocessing techniques	20
5.1	Structure of LSTM	22
6.1	Scores in Tabular format	30
6.2	Scores in Tabular format	31
6.3	Accuracy, precision, recall, and F1 score chart	32
6.4	Confusion Matrices	33
6.5	Confusion Matrices	34
6.6	Receiver operating characteristic Curves	36
6.7	Receiver operating characteristic Curves	37
6.8	Learning Curves comparison 1	38
6.9	Learning Curves comparison 2	39
6.10	Learning Curves comparison 3	39
7.1	Implementation of our model to a website	44

Chapter 1

Introduction

1.1 Inspiration for the Forecasting Model

Fake news refers to news that is intentionally fabricated and contains inaccuracies with the aim of misleading individuals into believing that it is genuine and factual. The term "fake news" gained widespread recognition among the general populace in 2016, during the U.S. election campaign, when Donald Trump employed it as a means of deflecting accusations levelled against him. The underlying cause of fabricated news can be attributed to either financial gain or political motives, with the intention of generating revenue or influencing public perception. In 2016, 62 per cent of U.S. people got their news through social media, contrary to one research [8]. By the conclusion of the presidential campaign, it is projected that over one million tweets would have been related to the false news "Pizzagate." In 2016, the Macquarie Dictionary named "fake news" the word of the year [8].

The dissemination of false information, commonly referred to as "fake news," has the potential to influence individuals' cognitive processes and sway their ideological leanings towards a particular viewpoint. Misinformation is frequently composed of a blend of factual and false information rather than solely consisting of falsehoods, as is commonly believed. There exists a well-known adage that states that falsehoods have the ability to spread rapidly, often reaching a wide audience before the veracity of the matter has been established. The rapid and unrestricted dissemination of information through digital technologies and social media has facilitated the widespread and expeditious propagation of false information, commonly referred to as "fake news," to a significantly larger audience. Frequently, online news platforms present fabricated news stories. The aforementioned news is fabricated and contains erroneous information with the intention of deceiving its audience. Nowadays, social media is inundated with this type of news. Individuals who produce fabricated news articles are primarily motivated by financial gain rather than a genuine concern for the subject matter they are reporting on. As an illustration, social media platforms such as Facebook can be used to attract traffic to websites, subsequently generating revenue through advertising. The writer profits from the number of people who click on the news as the news is designed as "clickbait."

Differentiating between fake news and authentic news can be a challenging task. The dissemination of false information, commonly referred to as "fake news," can

have significant ramifications when it transitions from the digital realm to the physical world [1]. In 2018, a case of false information regarding child abduction was disseminated widely through the messaging application WhatsApp in Mexico. Subsequently, a group of individuals, without verifying the veracity of the information, proceeded to set two men on fire who were believed to be involved in the alleged abduction. This incident highlights the dangers of misinformation and the potential for violent consequences resulting from the dissemination of false information. [12] An analogous occurrence transpired in India and Myanmar. Instances of lethal violence were incited in India, Myanmar, and Sri Lanka due to the dissemination of fabricated information on Facebook and WhatsApp [12]. The dissemination of false information, commonly referred to as "fake news," has been shown to have a detrimental impact on individuals' psychological well-being and can contribute to the proliferation of animosity within society. According to a study conducted by the Management and Resources Development Initiative (MRDI), there is a significant prevalence of fake news in Bangladesh. The survey revealed that rural areas have the highest rate of fake news experience at 66%, followed by urban areas at 62.3%, while metropolitan areas have the lowest rate at 52.5%. According to a recent study, approximately 50% of individuals do not make an effort to discern between factual information and subjective viewpoints when perusing news articles on the internet [24]. Our proposed methodology for creating a system of this nature involves the utilisation of machine learning methodologies for the purpose of detecting fabricated news content.

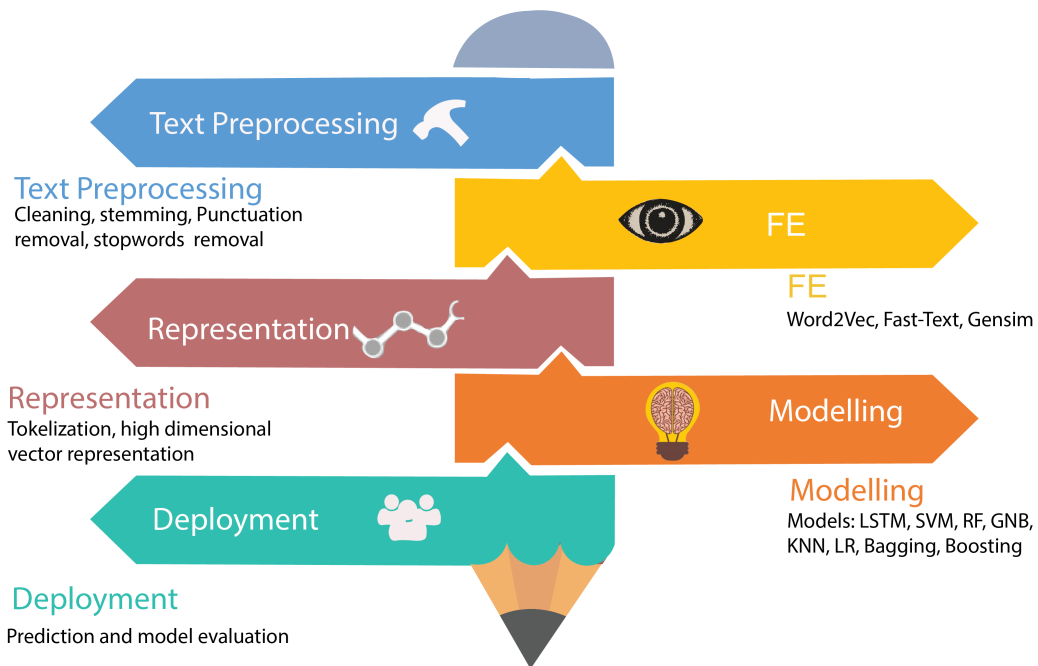


Figure 1.1: Roadmap

1.2 Research Problem

The traditional method of news consumption is rapidly changing. More individuals are joining social media daily, and news consumption through social media is be-

coming more popular than ever. On the other hand, the risk of deceiving people's attention and opinion is increasing at an alarming rate. According to [10], Click baits and fake news impede an audience's capacity to decoction useful information from internet services, mainly whenever news is essential for decision-making. Because of the dynamic structure of a modern business, the issue of false news has evolved beyond a business challenge but now requires substantial attention from cybersecurity experts. Any attempt to influence or inundate the Web with falsehoods or clickbait should be met with a resounding reaction. There are three main categories from a news and telecommunication standpoint.

There are three types of news headlines from a journalism and marketing standpoint: correct, vague, and deceptive [5]. An accurate headline has the same meaning as the substance of the news piece. Such headlines are not-clickbait. An ambiguous headline is one whose meaning is uncertain compared to the story's content. The omission of some crucial information is common in unclear headlines. The reader's curiosity is piqued by the absence of knowledge. It entices them to click. A misleading headline has a different connotation than the story's content. The distinctions can be subtle or glaring. Exaggeration, distortion, and other typical strategies are used to create a sensation. According to the concept of ambiguous news headlines, they frequently omit some critical aspects of phrases to elicit interest, which may be viewed visually without having to read the news body.

The available resources for identifying and verifying Bangla false information are inadequate. A limited number of research papers pertaining to the identification of fake news in the Bengali language have been discovered. Furthermore, the identification of Bangla news poses several challenges that are not present in the detection of fake news in English. To find the ambiguous headlines, we must first extract features from the headlines and tokenize words. However, those traits are primarily word-based, neglecting sentence structures and sequential information. Secondly, extract CSR features from sequential class rules (CSR). Then, the primary and CSR features are utilized for training a support vector machine (SVM) classifier [9]. According to [5], linguistic techniques e and assess the substance of deceptive communications to relate language patterns to deceit. The purpose of this linguistic method is to discover instances of predicted false indications inside the text's main content.. According to [3], name identification is essential towards the notion of social media credibility. Data dissemination throughout recent issues through mainstream technologies like microblogs necessitates techniques for discriminating between fake and 3 genuine content. In addition to the literature review, information and suspicious reference behaviour are used.

Our research project proposes a system in which users can input news articles and receive an output indicating the probability of the news being false. This system aims to reduce confusion and uncertainty surrounding the veracity of news articles.

The emergence of counterfeit news in the Bengali language has posed significant challenges in the age of digitalization. Despite the growing awareness of the adverse impacts of false information, there exists a dearth of efficacious instruments and approaches that are custom-made for identifying and countering fake news in the

Bangla language. The primary objective of this study is to fill a significant void by delineating the distinct linguistic attributes, trends, and obstacles linked to counterfeit news in the Bangla language. The research problem pertains to the creation of a comprehensive framework that utilizes machine learning algorithms, natural language processing techniques, and linguistic analysis to effectively differentiate authentic news from counterfeit news in the Bangla language. The objective of this study is to improve media literacy, advance information accuracy, and empower users, media organisations, and policymakers in their endeavours to counteract the dissemination of false information within the Bangla-speaking population through the resolution of research issues.

Chapter 2

Related Work

Here we will look at some machine learning and deep learning strategies that have shown promise in spotting fake news. What can and cannot be done to stop the spread of fake news and the methods by which it may be discovered in the news.

KLS Gogte Institute of Technology [14] researchers Chaitra K Hiramath and G. C Deshpande published a paper that utilized a system for detecting fake news based on classification. This system included Logistic regression (LR), Naive Bayes (NB), Support vector machine (SVM), Random forest (RF), and deep neural network (DNN). Comparative analyses of all machine learning approaches to the problem of identifying fake news were presented in the previous paper[14].

The article [15] authors Muhammad and Yousaf, Suhail and Ahmad, and Muhammad Ovais proposed a machine learning assembly technique for automatically categorizing the news and their piece. They also investigated many linguistic characteristics that might be utilized to spot bogus news from the actual one. They trained a mixture of distinct machine-learning algorithms using various ensemble approaches and then assessed their performance on four real-world datasets. The results of their experiments showed that the suggested ensemble learner strategy outperformed individual learners by a significant margin. The framework used several steps, such as collecting multiple datasets from different domains, pre-processing the data, classifying articles, 70/30 training/testing Ensembles, studying bagging, boosting, and voting classifier ensembles, and studying bagging, boosting, and voting classifier ensembles. Accuracy, precision, recall, and F1 score assess model performance. The authors analyze four public datasets to demonstrate their superiority. The highest accuracy was achieved from SVM, and logistic regression, which is 92%, and the lowest accuracy was achieved by 27%, which is by Lstm while combining metadata elements with texts and while they have 88.6% accuracy on MLP. Research and performance assessments employ their method[15].

Despite the fact that a significant amount of work has been done for the English language and several other languages to create techniques and to boost the efficiency of fake news detection, research on the identification of fake news within the Bengali language remains in the early stages of development. Tohabar [22], Md Yasmi, Nasrah, Nahiyah, Samir, and Asif Mohammed authors detected bangla fake news through the study's efforts to create a mechanism for detecting fraudulent Bengali-language news reports utilizing a range of datasets and machine learning approaches. For analysis, the following data sets were used: Authentic-7K Dataset of False News Stories: The collection contained 7,202 news pieces, all classified as true or false. A

total of 3,964 reports were confirmed to be true, while 3,238 were shown to be fake. Article identifier, domain, date, category, source, relation, headline, content, label, and so on were only a few of the columns in the dataset[22].

This dataset, dubbed LabeledFake-1K, contained 1,299 items of news that had been categorized as clickbait[22], satire, or false news. Clickbait news items numbered 82, parody news pieces 1,136, and false news articles 81. This data collection includes an F-type column and features such as LabeledAuthentic-7K. In this work, the classifiers Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) were examined. The datasets were divided in an 80:20 ratio between training and testing sets, and the algorithms' performance was measured using accuracy, precision, recall, and F1 score. In their proposed model, they had the highest accuracy in SVM when they added sentiment recall and got higher accuracy, and the F1 score had lower accuracy than they showed.

In paper [5], the authors propose to make a fake news detection tool that follows a veracity assessment method and a hybrid approach. Two machine learning techniques are used in the methods, which are the linguistic approach and the network approach. In the linguistic approach, texts can be checked for signs of fraud using a Bag of Words, Deep Syntax, Semantic Analysis, Rhetorical Structure, Discourse Analysis, and Classifiers. Bag of Words[5] detects deception signals by analyzing the frequency of individual words or multiword phrases (n-grams), while Deep Syntax uses probability context-free free grammar (PCFG). To determine validity, Semantic Analysis compares human experiences to content profiles obtained from similar data. Classifiers employ [2] word and category frequencies to predict deceit based on numerical grouping and distances, whereas Rhetorical Structure and Discourse Analysis make use of rhetorical relations to suggest dishonesty. Knowledge networks use semantic proximity and network relationships to assess truth. The hybrid approach combines linguistic and network approaches to improve accuracy[5]. In their proposed methods, they have success on SVM classifiers performing 86% as well as human judges in identifying spam with false, negative opinions.

In the paper[20], the authors proposed that their research delves into cutting-edge methods for spotting false news and discusses relevant datasets and natural language processing (NLP) methods. In the article[20], they give a thorough introduction to the methods that make use of deep learning. The challenges they face and their research directions are as they mentioned. Although research on the detection of false news has been done, there is always an opportunity for improvement and inquiry. Although DL-based approaches offer more accuracy, there is still a need to improve their acceptability. To boost performance, researchers should concentrate on feature and classifier selection. It is necessary to investigate features like user behavior, user profiles, and social network behavior. Since there isn't much propagation-based research in this field, it's important to manage meta-data and other information properly. Their Combining DL and ML techniques to build an ensemble model yields superior results. GRU models perform better than LSTMs, so a combination of GRU and CNNs is suggested for the best outcome[20].

Ranjan and Aayush[13] authors stated in their research that In this study, the application of machine learning methodologies is employed to ascertain the veracity of news articles by analysing their textual content and user-generated responses. Algorithms like Support Vector Machine, Passive Aggressive Classifier, Multinomial Nave Bayes, Logistic Regression, and Stochastic Gradient Classifier are trained using

frequency-based features. The findings of an experiment that tried to find instances of Fake-related phrases in answer texts were encouraging. The challenging thing was most crucial information presented here is that social media users' opinions on postings may be utilized to assess stories' credibility and that Linear Support Vector Machines using Tf-Idf vectors achieve the greatest classification accuracy (93.2%), sensitivity (92%), and ROC AUC score (97%). It is also possible to apply additional linguistically based criteria to comments in order to ascertain the validity of news. Because of its high ROC AUC score (97%), high sensitivity (92%), and high classification accuracy (93.2%), this approach was used as a foundational component in Fake news identification, as their result showed.

In [19], it has been discovered that the researchers, in order to identify and categorize bogus articles, this study provide a methodology for doing so utilizing supervised machine learning algorithms and feature selection techniques. Their research examines the two phases—characterization and disclosure—of the process of spotting false news. Several supervised learning techniques currently in use for detection are discussed. The accuracy of Naive Bayes is 96.08%, that of a Neural Network is 76.08%, and that of a Support Vector Machine (SVM) is 76.08%. The most salient facts are that using KNN and random forests increased false message detection accuracy by up to 8% and that using N-gram analysis with a unigram and linear SVM algorithm resulted in the best accuracy. It was also suggested that numerical statistical values be included as characteristics using POS textual analysis in order to enhance precession findings.[19]

Authors [7] Ruchansky, Natali and Seo, Sungyong and Liu, and Yan, in their research paper, mentioned that their study recommends a strategy that combines three factors for better prediction. The system is made up of three components: integration, scoring, and capture. The system's first component tracks reader interactions with a particular piece of material, and its second component uses that information to infer a source attribute. Real-world data analysis demonstrates that CSI performs better than cutting-edge models and can consistently extract useful latent representations of users and articles.

Two real-world datasets, Twitter and Weibo were used in the testing to compare the proposed CSI model against cutting-edge models. The results showed that CI-t surpassed CI in terms of accuracy and F-score [10] by more than 1%, whereas CSI outperformed other models in terms of accuracy and F-score. This suggests that although linguistic traits could have some temporal aspects, it may be possible to tell the difference between true news and false news based on the frequency and distribution of encounters [7]s. In their success, CSI performs better than all variations and comparison models. We can see that the overall results from GRU-2 rise by 4.3% when user features are included. Together, these findings demonstrate how CSI effectively captures and makes use of all three characteristics—text, response, and source—to classify fake news

Chapter 3

Methodology and Requirement Analysis

3.1 Data Collection

The initial step of our methodology involved procuring a Bangla language corpus comprising genuine and contrived news articles. The dataset plays a crucial role in machine learning and deep learning systems, as it directly impacts the system's quality, accuracy, and other relevant metrics. It is imperative that the dataset exhibits diversity, balance, and accuracy in its representation of the target population. The objective of our Bangla Fake News Detection system was to construct a comprehensive database of news articles spanning a range of categories, including political, social, national, international, miscellaneous, technological, criminal, lifestyle, and economic news.

The data was gathered through a diverse range of techniques, encompassing web scanning, API queries, and manual data entry. The data was gathered from diverse sources, including online forums, news outlets, and social networking websites. We conducted a search for news articles by utilizing hashtags and keywords that are linked to current events transpiring in Bangladesh. Furthermore, a compilation of news articles sourced from reputable news websites and message forums, including but not limited to bdnews24.com, prothomalo.com, and somoynews.tv, was conducted. However, the task of discerning fraudulent news articles can be arduous. Bangla is a language that is widely spoken by a significant populace across several countries, such as Bangladesh, India, and certain areas of Myanmar. Consequently, the task of detecting and revealing fraudulent sources of Bangla news and reports poses a challenge.

These regions disseminate misinformation due to inadequate comprehension and education regarding the ramifications of disseminating false news. A significant number of people do not possess the necessary knowledge to distinguish between authentic and fabricated news, and inadvertently aid in the dissemination of misinformation. Moreover, the ubiquity of social media platforms and the convenience of virtual data transmission have expedited the swift dissemination of erroneous data. These websites are often characterized by insufficient oversight and fact-checking, which increases the likelihood of erroneous data going unnoticed.

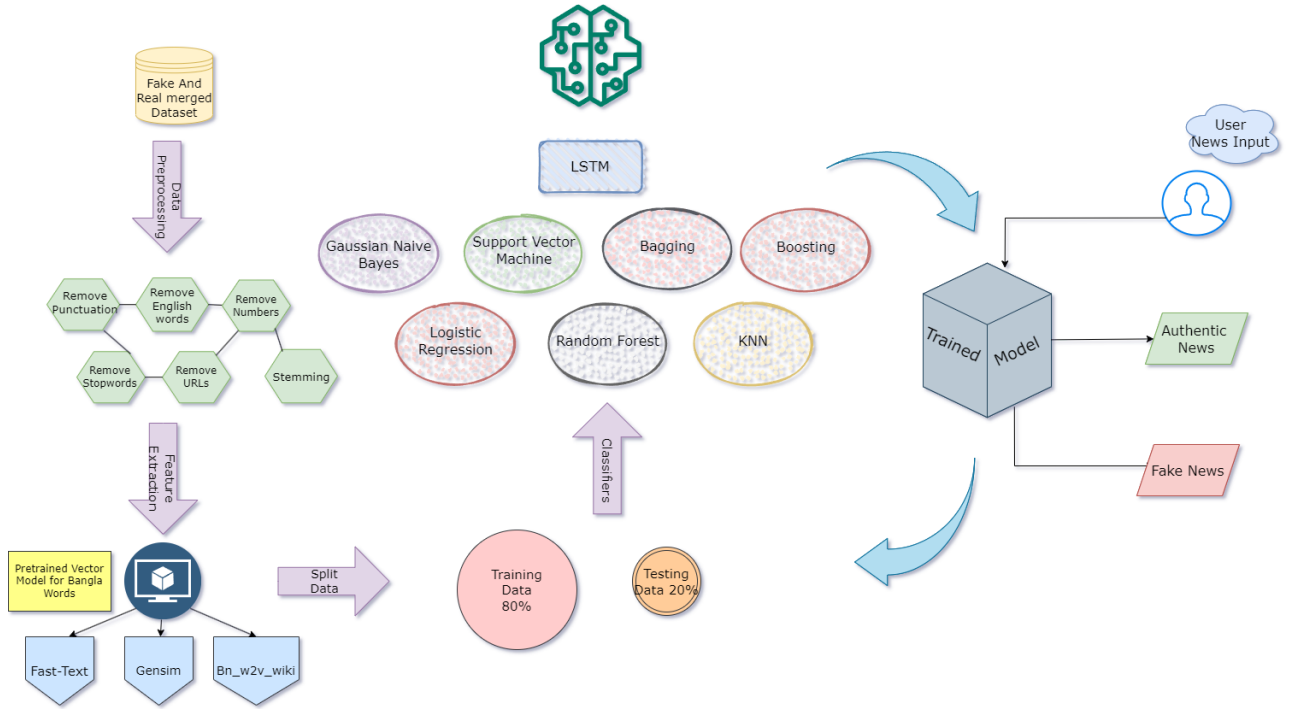


Figure 3.1: Architecture of our system

A meticulous methodology was implemented for the labeling of data in order to guarantee the accuracy and consistency of the dataset. The members of the thesis committee underwent a rigorous training process that encompassed guidelines for data labeling, examples of authentic and counterfeit articles, and a training session. The news articles were independently labeled by the members and any discrepancies were resolved through discussion and agreement.

3.2 Data Preprocessing

The data preprocessing phase is essential to any system based on machine learning or deep learning. In this stage, the unprocessed data is transformed into an arrangement that is appropriate for evaluation and modelling. Data cleansing, tokenization, stop word removal, stemming, and feature extraction are all included in the data preprocessing phase. In the context of our Bengali Fake News Detection system, the subsequent data preprocessing duties were performed:

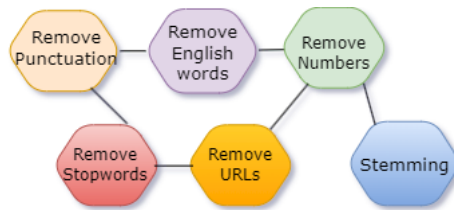


Figure 3.2: Preprocessing stages

3.2.1 Remove Punctuations

During the process of data cleansing, one of the many tasks involved is the elimination of HTML elements, punctuation marks, numerical values, and stop words from each column of our datasets. There exist several Python libraries designed for the English language that are capable of effectively removing such redundancies from the given dataset. Nonetheless, the Bangla language has a limited number of tools available for classifiers to eradicate these extraneous components. There exist several libraries capable of executing said tasks, however, their efficacy is suboptimal and their utilization presents difficulties. The insufficiency of the functions and exception handling descriptions provided by the proprietor’s documentation is apparent. The BNLN library [21] was utilized to remove punctuation, numerals, and English words from the text.

The BNLN Langtools library has been specifically developed for the purpose of Bengali natural language processing. In addition to various other functionalities, the system provides tokenization, stemming, part-of-speech tagging, identification of named entities, and analysis of sentiment. The BNLN Langtools are widely utilized by the Bengali natural language processing (NLP) community for the purposes of studying, generating, and analyzing linguistic data pertaining to the Bengali language. The HTML elements present in the media reports were deemed unnecessary to our analysis and were therefore removed. Furthermore, the exclusion of punctuation and special characters from the news articles was carried out as they were deemed irrelevant to our analysis. Furthermore, we excluded all numerical data from the news articles as it was deemed extraneous to our analysis.

3.2.2 Remove Stopwords

In the Bangla language, stopwords refer to frequently used words that do not significantly contribute to the overall meaning of a sentence. The predominant portion of these lexical items comprises prepositions, conjunctions, articles, and pronouns. Some examples of Bangla stopwords include "একটা", "একটি", "এমন", "ইত্যাদি", "ও", "তার", "তারা", "সে", "সেই", and "যার".

The identification and removal of stopwords is a crucial step in the analysis of Bangla textual data. The reason for this is that the incorporation of such terms in the analysis may lead to imprecise outcomes and a rise in computational duration. The data collected underwent stopword elimination through the utilization of BLTK[26] language tools. The BLTK (Bangla Language Toolkit) is a specialized library that has been developed with the purpose of facilitating the processing of the Bangla language. The platform provides a diverse range of resources and tools that facilitate natural language processing, text classification, and sentiment analysis of Bangla text.

The library houses linguistic resources pertaining to the Bangla language, including stopwords, stemming, and POS labeling. These resources can be effectively employed for the purpose of constructing language models and developing applications that necessitate text analysis or generation.

The BLTK [26] offers a comprehensive feature for eliminating stopwords in the Bangla language. The level of stopwords removal has been set to moderate to ensure that similar words to stopwords are not lost, and irrelevant stopwords are not retained, thereby maintaining the relevance of our classifiers.

3.2.3 Stemming

The method to reduce conjugated or derived phrases to their base or root shape is called stemming. It is a crucial step in natural language processing because it simplifies the text, making it simpler to understand and process.

For the Bangla language, stemming is an important aspect as it is a highly inflected language with a rich morphology. For example, the word "কাটা" (kata) meaning "to cut" can be stemmed to "কাট" (kat) by removing the suffix "া" (a).. Numerous efforts have been made to create stemming algorithms for the Bengali language, and diverse methodologies have been suggested.

A widely used method is the rule-based stemmer, which employs a predetermined set of rules to recognize the base form of a given word . An alternative methodology is the implementation of a statistical stemmer, which leverages machine learning algorithms to deduce the base form of a word by analyzing its frequency of appearance within a given corpus. The author requires a citation for the information provided. Notwithstanding, the creation of efficient stemming algorithms for the Bangla language remains a formidable undertaking, owing to the intricacy of the language and the dearth of superior linguistic resources. However, through ongoing research and development, it is reasonable to anticipate improvements.

In the course of our study, we have experimented with various stemmers designed for the Bangla language. Nonetheless, we have encountered numerous challenges in utilizing them due to the library's intermittent loading caused by unidentified inconsistencies. Initially, the 'bengali-stemmer' developed by Rafi Kamal and hosted on the Github repository owned by banglakit [4] was employed. The outcome obtained was not in accordance with our intended objective. The UrgaStemmer tool from the BLTK[23] library in Python has been utilized in our study. Consequently, we encountered compatibility issues with the stemmer. Ultimately, the 'bangla-stemmer 1.0' developed by Mezbaul Islam Protick [25] was utilized as it demonstrated superior performance in comparison to other stemmers that were employed.

3.3 Feature Extraction

The technique of feature extraction holds significant importance within the domain of natural language processing. Its primary objective is to transform textual data into a numeric representation, which can be effectively utilized by machine learning or deep learning algorithms. The process of feature extraction encompasses the conversion of textual data into a collection of features that accurately capture the semantic essence of the text. Our Bangla Fake News Detection system underwent experimentation with various feature extraction techniques, such as Word2Vec, FastText, and GloVe. Utilizing pre-existing word embeddings offers the benefit of reducing the amount of resources and time required for training self-generated embeddings from the ground up. Furthermore, pre-existing embeddings are frequently trained on copious quantities of textual data, thereby resulting in more accurate embeddings in comparison to those trained on limited datasets.

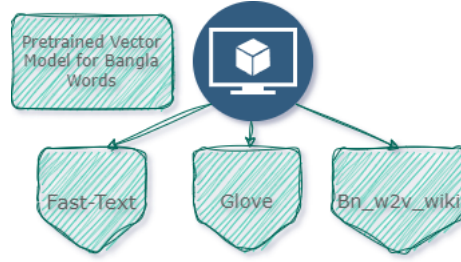


Figure 3.3: Feature extraction methods

3.3.1 Word2vec

The Word2Vec technique is a commonly utilized method for extracting features in the field of natural language processing. The method utilized involves the representation of words as vectors with high density in a multi-dimensional space, where the spatial arrangement of each word vector corresponds to its semantic connotation. The Word2Vec methodology involves the utilization of a neural network that is trained on a vast corpus of textual data to forecast the context of individual words. The word vectors that ensue are indicative of the semantic content of words in relation to their contextual surroundings.

The utilized Bengali Word2Vec Model, namely 'bn_w2v_model' [6], is a pre-trained model by Firoj Alam et al. The present model has been constructed for utilization with the BNLNLP [21] package, and it should be noted that the dimensionality of the word embeddings employed is 300. This entails the selection of a term from the dataset, followed by its conversion into a 300-dimensional vector, which can subsequently be utilized for the training of our classifier models. The process of vectorization is employed due to the inability of machine learning or deep learning models to comprehend textual data. The data needs to be quantified in numerical form. Thus, the Word2Vec model is introduced to occupy the position.

Prior to integrating Word2Vec into our Bangla Fake News Detection system, we initially conducted preprocessing procedures on the Bangla news articles, as previously outlined. Subsequently, a Word2Vec algorithm was employed to train on a vast Bangla textual dataset, thereby producing word embeddings. The model that underwent training produced compact vector representations for every Bangla term present in the collection of data, which were subsequently employed as attributes for our machine learning and deep learning methodologies.

3.3.2 Bangla FastText Model

FastText is a variation of the Word2Vec model which was developed by Facebook. The operational mechanism involves the training of a neural network to forecast character n-grams, as opposed to sentences. The connections that ensue possess the capability to encapsulate not just the semantic essence of the words themselves but also the underlying sub-word configuration, thereby rendering it more efficacious in managing infrequent and unfamiliar words.

The pre-trained FastText model 'bangla-fasttext' developed by Sagor Sarkar [21] has been constructed for the Bangla language. This has also been developed for the BNLNLP software package. The Fasttext model was trained on a corpus of 20 million words, resulting in a vocabulary size of 1,171,011. The model was trained

for 50 epochs and utilized an embedding dimension of 300. The pre-trained model is utilized to transform each word in the corpus into a 300-dimensional vector. Incorporating FastText into our Bangla Fake News Detection system was executed through a comparable methodology as that of Word2Vec. The Bangla media reports were preprocessed prior to training a FastText model on a substantial corpus of Bangla text data. The embeddings obtained were able to effectively capture both the semantic and sub-word structural aspects of the Bangla language, which were subsequently utilized as functionality for our machine learning and deep learning models.

3.3.3 Gensim

Gensim is a widely utilized library for feature extraction in natural language processing endeavors, alongside Word2Vec and FastText. The tool offers a proficient and user-friendly execution of Word2Vec and other embedding methodologies. The Gensim tool was employed in our Bangla Fake News Detection system for obtaining word embeddings from the article data. The gensim library offers a module called KeyedVectors, which facilitates the utilization of pre-existing word embeddings by users. Word embeddings refer to of 300 dimensional vector representations of words that accurately convey their semantic and contextual meaning. The KeyedVectors module provides the functionality to load pre-existing word embeddings, such as GloVe or Word2Vec, and execute diverse operations on them, such as identifying analogous words, computing word resemblance and identifying words that are identical in meaning.

In order to perform feature extraction using Gensim, it was necessary to create a corpus consisting of a set of data. In the present study, it was observed that every individual record denoted a pre-processed news article in the Bengali language. The corpus serves as the input for the Gensim model, which is utilized for both learning and developing word embeddings. The aforementioned word embeddings refer to compact numerical vectors of 300 dimension that encapsulate the semantic essence of words. The word embedding for a given word can be acquired by obtaining it through the trained model.

Integration with Machine Learning/Deep Learning Models To utilize word embeddings as features to feed our machine learning or deep learning algorithms, we converted every news document into a numerical form which is 300 dimensional vector representation through the use of said word embeddings. A prevalent methodology involves the computation of the arithmetic or weighted mean of the word embeddings found within a document, which results in the acquisition of an illustration at the content level.

As an illustration, the word embeddings for every word in a news article were obtained from the Gensim-trained Word2Vec model. Subsequently, the mean of said word embeddings was calculated, yielding a constant numerical vector 300 dimension that denotes the entirety of the news article.

The vector representations derived from the Gensim Word2Vec model were utilized as feature inputs in our machine learning or deep learning algorithms. The aforementioned features may be utilized in the training and assessment of models designed for the purpose of detecting fabricated news in the Bengali language.

```

array([ 2.69087e-01, -2.66610e-02,  7.58100e-02, -6.78200e-03,
       -5.35660e-02, -8.79370e-02, -2.07028e-01, -6.80460e-02,
        7.28310e-02,  1.89679e-01,  1.85252e-01, -5.52990e-02,
       -6.02450e-02,  4.87610e-02, -1.46224e-01,  5.00720e-02,
        1.06839e-01,  1.90304e-01, -1.22210e-02,  1.94980e-01,
        5.05480e-02,  2.45947e-01,  3.38357e-01, -2.12400e-03,
       -1.95300e-02, -1.30105e-01, -8.13180e-02, -5.72620e-02,
        1.44294e-01,  1.77986e-01,  2.68190e-02, -2.26740e-01,
       -1.64151e-01,  2.52168e-01, -3.06169e-01, -1.23330e-02,
        1.39819e-01,  8.61550e-02, -3.71860e-02,  1.63239e-01,
       -9.39760e-02,  3.61890e-02, -1.30990e-02,  4.87070e-02,
       -1.41940e-01,  1.02602e-01, -2.25119e-01,  5.63760e-02,
       -1.03305e-01, -1.43948e-01, -2.16863e-01,  2.58318e-01,
       -5.07180e-02,  1.12299e-01, -1.22830e-01, -1.23497e-01,
       -8.87480e-02,  2.75765e-01,  2.27037e-01, -9.75960e-02,
        7.39330e-02, -6.94610e-02,  2.61699e-01,  2.80349e-01,
       -2.63300e-02, -1.07007e-01, -1.45988e-01,  2.31703e-01,
       -2.88790e-01, -1.02423e-01, -2.06376e-01,  2.50020e-02,
        1.22547e-01, -1.61182e-01, -2.61320e-02, -5.84150e-02,
        5.23010e-02, -3.42524e-01,  3.71900e-02,  6.91300e-02,
        2.60230e-02, -6.66650e-02, -5.70690e-01, -1.24667e-01,
        1.90898e-01, -1.66774e-01, -7.15010e-02, -1.48678e-01,

```

Figure 3.4: Glimpse of a 300d vector

3.4 Smote

The dataset was analysed to determine the prevalence of fake news instances in comparison to real news instances. It was found that fake news instances were significantly fewer than real news instances. To address this class imbalance, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm was employed as a potential solution. The SMOTE technique is a widely used approach for generating synthetic samples of the minority class, specifically in the context of fake news, in order to achieve a balanced dataset.

The issue of class imbalance is tackled by SMOTE through the generation of artificial instances belonging to the underrepresented class. The methodology involves the identification of minority class instances and the generation of synthetic instances by connecting neighbouring instances along line segments. The SMOTE algorithm employs a hybrid approach of oversampling at random and interpolation methodologies to produce artificial instances.

The technique of SMOTE is employed to augment the representation of the minority class in the dataset by generating synthetic samples. This approach is utilised to increase the presence of fake news instances in the dataset. The implementation of balancing techniques is crucial to mitigate the impact of class imbalance, thereby ensuring a fair and unbiased allocation of instances for classifier training.

Following the balancing of the dataset through SMOTE, a classification model was trained on the enhanced dataset. By utilising both synthetic and real instances of the minority class, the model can enhance its capacity to accurately classify fake news. Following the training phase, the model’s performance was assessed on the test set using relevant metrics, including accuracy, precision, recall, and F1 score.

Synthetic Minority Oversampling Technique

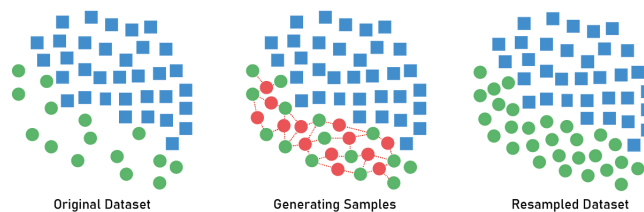


Figure 3.5: Smote Technique

src: <https://emilia-orellana44.medium.com/smote-2acd5dd09948>

Chapter 4

A comprehensive look at the dataset

A credible and extensive dataset is critical in constructing a practical artificial intelligence (AI) model for detecting false news. A good dataset is essential because it can give various trustworthy tagged instances that effectively depict the many features of false statements. With such a dataset, the AI model may learn and detect the subtle patterns, language signals, and contextual information that differentiate fake news from authentic material. A high-quality dataset guarantees that the model is trained solidly, increasing its capacity to generalize and forecast accurately when exposed to previously unreported instances of false news. Furthermore, a good dataset helps assess and compare different AI models, allowing researchers and developers to quantify the usefulness and efficiency of their detection systems. Finally, an immaculate dataset serves as the foundation for creating steadfast AI models, boosting the battle against the spread of fake news and promoting information integrity in the digital world.

The dataset used for training our models combines main and secondary datasets. To begin, the authors of [16] did a wonderful job creating the dataset, which contains roughly 50,000 news items, including authentic news, labeled authentic news, fake news, and labeled fake news. We would like to take this opportunity to thank Hosain et al. (2020) for their outstanding contribution. This dataset will be a valuable resource for building tools to counteract the spread of fake news as well as research in languages with limited resources. In this dataset, 22 of Bangladesh's most reliable mainstream news publications compiled a list of credible news sources. This dataset contains 48k authentic news, 1k false news, 7k genuine news with labels and 1k fake news with labels. The writers have assigned the value '1' to legitimate news and '0' to fake news. In the labeled news, the authors specified the sorts of datasets that have misleading or false contexts, containing incorrect information or content that might mislead viewers.

The relevance of a balanced dataset in constructing an AI model for false news detection cannot be emphasized, given the abundance of legitimate news vs fraudulent news. To address this issue, our contribution in this sector required a tedious manual collection of fake news stories taken from infamous news sites recognized for their significant effect on consumers. Our primary goal was to increase the number

of instances of fake news, which are noticeably sparse in comparison to their true counterparts. To assure the correctness and relevancy of the gathered false news stories, thorough curation techniques were used, followed by categorizing them to fit with the category labels of our secondary datasets. This methodical technique enabled us to build a large corpus of around 1500 fake news pieces, each tagged properly. These fake news incidents were then easily combined with the secondary dataset, yielding a much larger and more more diverse dataset ready for further processing and analysis.

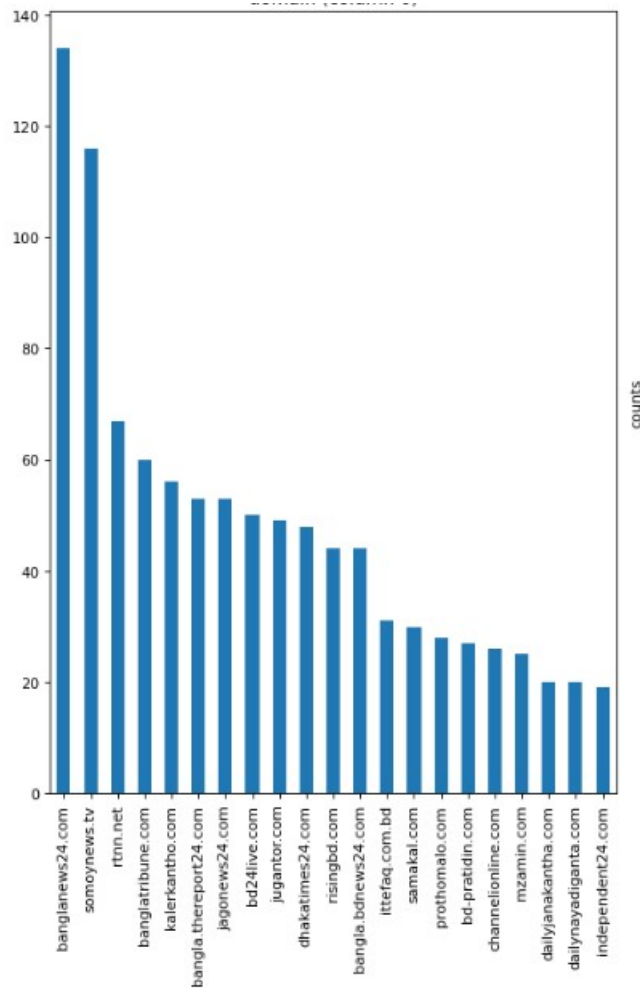


Figure 4.1: Authentic news sources count.

Source:

<https://www.kaggle.com/code/kerneler/starter-banfakenews-4d21e637-c/notebook>

Clickbait refers to using provocative headlines in news items to attract readers and increase website clicks for the publisher. In contrast, Satire/Parody refers to humorous and entertainment-oriented news reporting. Based on an analysis of the dataset, it has been determined that banglanews24.com , somoynews.tv , and rtnn.net are the most reliable news sites. In contrast, sites like, channeldhaka.news, eariki.com, banglaviralnews.com are some major notorious sites. Additionally, the categories "National", "International" and "Sports" are given the highest emphasis in real news. Contrarily, "Miscellaneous" and "Politics" are prioritised in Fake News. A

visualizaaiion has been provided below for clear understanding:

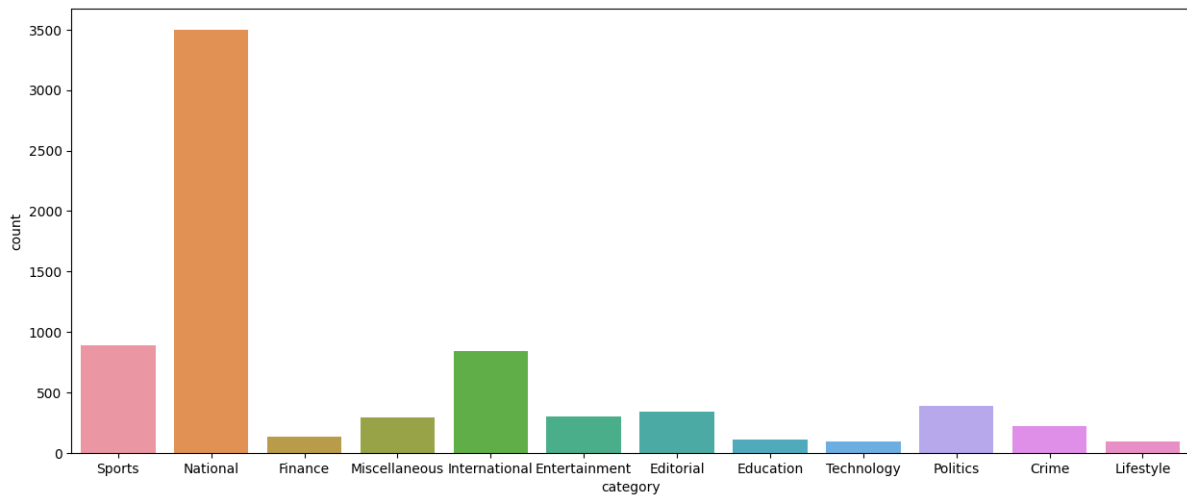


Figure 4.2: Authentic news categories

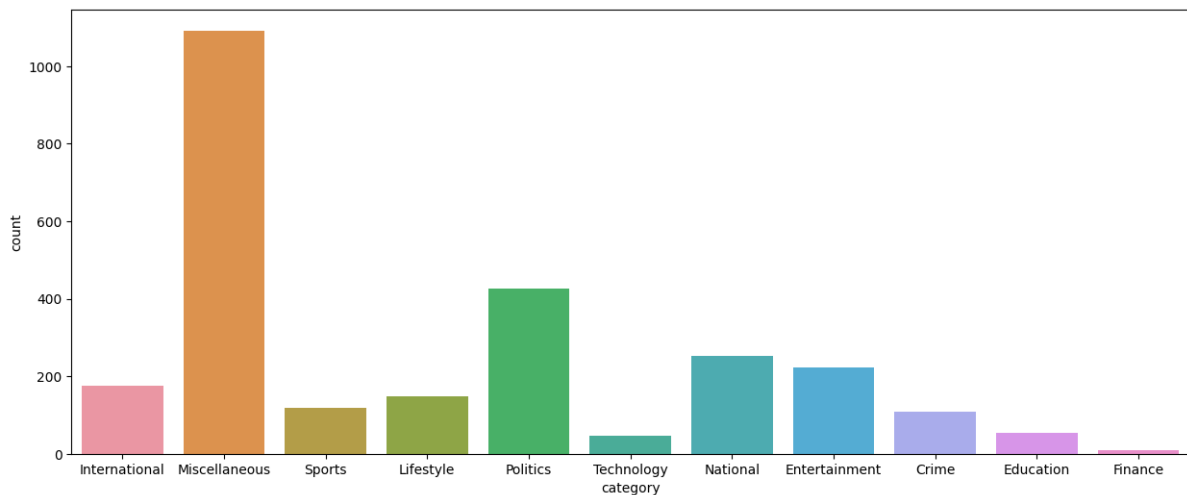


Figure 4.3: Fake news categories

The magnitude of each word in these word clouds corresponds to their frequency or significance in a given context. In a word cloud, words are presented in various sizes and orientations to create an aesthetically pleasing pattern. This technique is frequently employed to summarise and illustrate the most prominent or frequently used words in a text corpus. By separating a large amount of text into a concise visual display, word clouds offer a concise overview and convey the essence of the underlying content. They enable readers to quickly comprehend the major themes, trends, or emphasis of the text. The greater the magnitude of a word in the cloud, the greater its frequency or importance in the original text. This intuitive representation aids in the identification of key terms, the extraction of meaningful insights, and the facilitation of further exploration or analysis of textual data. Our word cloud of real and fake datasets aids us in comprehending the influence of certain words on the reader's perception.



Figure 4.4: Authentic news word cloud



Figure 4.5: Fake news word cloud

Now that our dataset has been obtained, it is time to clean it. Cleaning and preprocessing datasets are essential for guaranteeing the quality and dependability of data used for study and modelling. Cleaning involves identifying and resolving various data problems, including missing values, anomalies, duplicate data fields, and inconsistencies. By resolving these concerns, we can improve the precision and dependability of our analysis outcomes. In order to make data appropriate for analysis, preprocessing entails transforming and standardizing it. This may involve duties such as feature scaling, normalization, encoding categorical variables, and textual data manipulation using techniques such as tokenization and stemming. Not only do proper dataset cleansing and preprocessing improve data quality, but they also help reduce bias, enhance model performance, and improve interpretability. It ensures the data is in a reliable and usable format, allowing for efficient analysis, modelling, and decision-making. By investing time and effort in these crucial stages, researchers can ensure the validity and reliability of their findings, resulting in more robust and insightful conclusions.

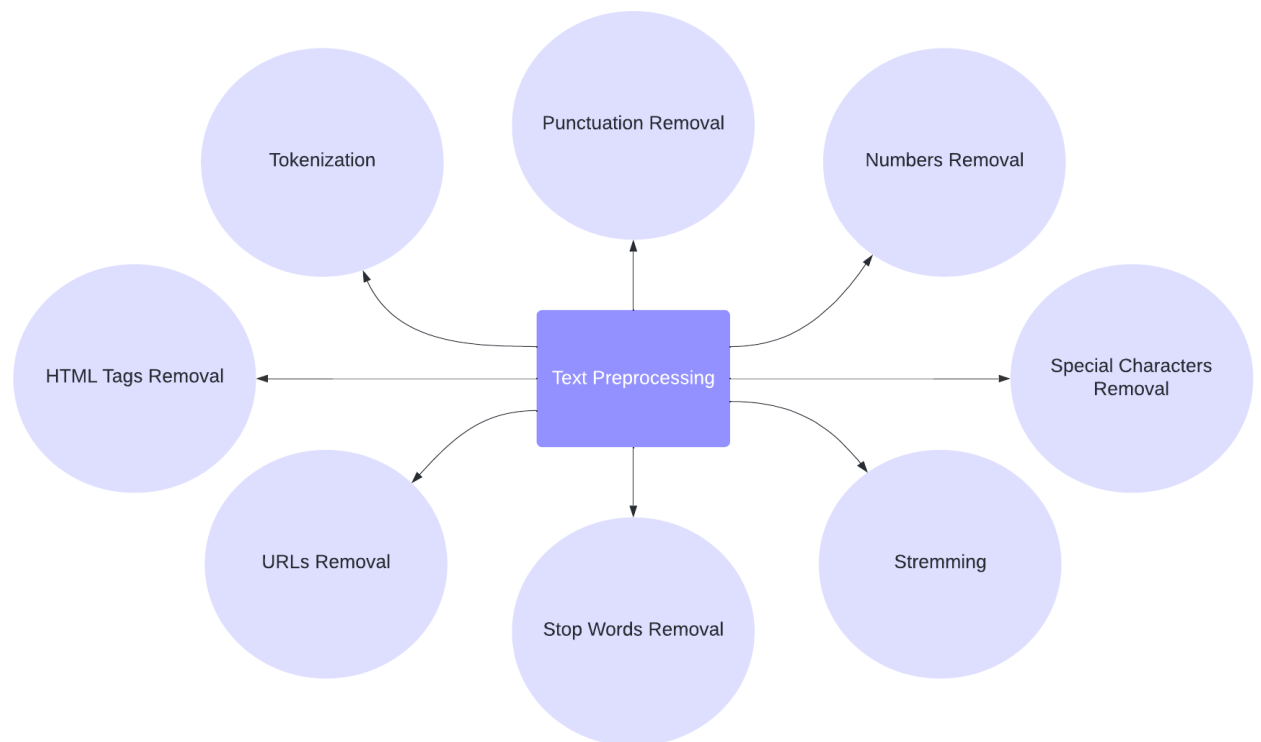


Figure 4.6: Data Preprocessing techniques

Chapter 5

AI Models and Implementations

During the feature extraction stage of the fake news detection process, the pre-processed text is transformed into numerical representations. The numerical representations effectively represent both the semantic significance of the words and contextual information. This enables the deep learning model to effectively process the input and learn information from it.

In our present study, we selected word embeddings as the principal technique for feature extraction, given its convenience in natural language processing (NLP) applications, including the identification of fake news, among other feature extraction methods. Furthermore, given the constraints of feature extraction techniques in relation to the Bengali language, word embeddings represent the optimal choice for our research objectives.

The utilization of word embeddings is a crucial methodology in the representation of words as dense numerical vectors. Word embeddings are capable of capturing semantic associations among words, thereby enabling the model to comprehend similarities, dissimilarities, and contextual associations.

Word embeddings offer distributed representations for lexical items, whereby each term is denoted by a vector of continuous numerical values as opposed to a discrete symbol. This depiction encapsulates the concept that words possessing analogous meanings ought to possess comparable vector representations, thereby empowering the model to exploit semantic associations. The Word2Vec model is a widely used technique for generating word embeddings, which involves the acquisition of word representations from extensive collections of textual data. The text presents a pair of primary methodologies, namely the Continuous Bag-of-Words (CBOW) and Skip-gram. The Continuous Bag of Words (CBOW) model endeavors to forecast a specific word by taking into account its surrounding context. Conversely, the Skip-gram model aims to predict the context words by utilizing a given target word. Both methodologies aim to optimize the embedding vectors with the objective of maximizing the likelihood of accurately predicting the intended words. The utilization of Word2Vec embeddings has become prevalent owing to their straightforwardness, effectiveness, and capacity to apprehend syntactic and semantic relationships.

The current study employs the `bn_w2v_model` is a highly proficient Bengali Word2Vec model, which has demonstrated efficacy throughout the entire process. This model is compared to other cutting-edge feature extraction models like GloVe and FastText. The model's vocabulary size is 436126, and its vectors are 300-dimensional.

5.1 LSTM

The development of an LSTM (Long Short-Term Memory) model to detect fake news includes the design of a deep learning framework that takes advantage of the capabilities of LSTM layers to capture sequential dependencies present in textual data effectively. The Long Short-Term Memory (LSTM) is a Recurrent Neural Network (RNN) variant that has been developed to effectively process sequential data, including but not limited to speech, text, and time series. Long Short-Term Memory (LSTM) networks exhibit the ability to acquire long-term dependencies in sequential data, rendering them highly appropriate for various applications, including but not limited to language translation, recognition of speech, and time series prediction.

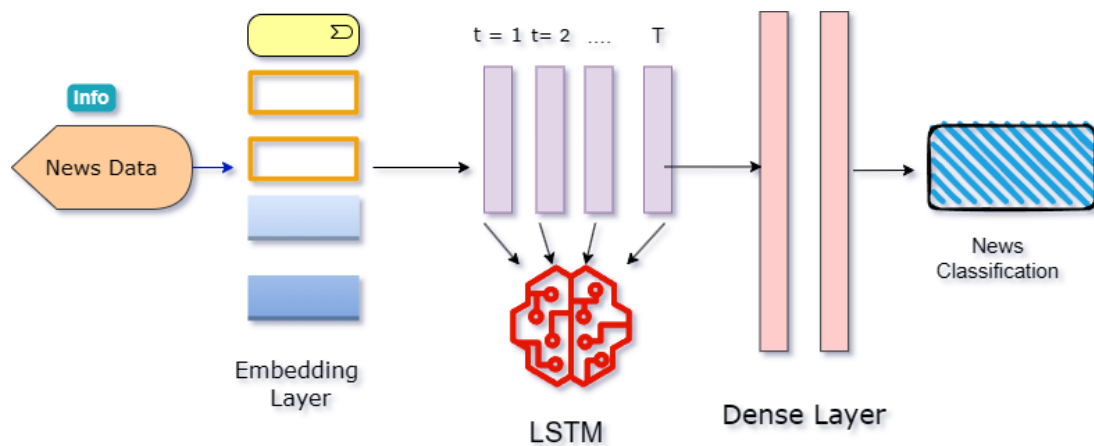


Figure 5.1: Structure of LSTM

Subsequently, the model incorporates the embedding layer. The task at hand involves the transformation of word indices into compact, fixed-size vectors. The embedding layer encompasses various parameters, including `vocab_size`, which denotes the overall count of distinct words present in the vocabulary. The length of the input serves as a determining factor for the maximum allowable length of a news item that is to be considered. Sequences that exceed the maximum length will be subject to truncation, while sequences that fall short will be subject to padding. The `weights` parameter maintains a significant role in this context as it takes the value of the `embedding_vectors`. The embedding layer is initialized with pre-existing word embeddings. The variable `embedding_vectors` is required to be a matrix in which each individual row represents the embedding vector associated with a particular word. When the `trainable` parameter is set to 'True', the embedding layer's weights can be modified during training, which facilitates the acquisition of task-specific embeddings by the model.

The sequential model incorporates an LSTM layer. The number of LSTM units or memory cells was configured to 128. Augmenting the number of units enhances the model's capability to apprehend intricate patterns, however, it also amplifies the computational expenditure.

The model is augmented with a dense layer comprising only one neuron. The output is subjected to the sigmoid activation function, resulting in a probability score that ranges from 0 to 1. This is appropriate for tasks involving binary classification,

such as the identification of fabricated news. Ultimately, the model is assembled utilizing the designated optimizer, loss function, and metrics. The loss function employed in our binary classification problem is binary cross-entropy, which is a widely utilized approach. The model is subsequently highlighted for its emphasis on specifying accuracy as the evaluation metric to be utilized during both the training and evaluation processes.

Train and Test

In order to facilitate training, the input features or independent variables are denoted as 'X' while the corresponding labels or dependent variables are denoted as 'y'. The tokenizer class from the Keras Library is utilized in the preparation of our 'X'. The system generates an internal vocabulary by utilizing the input textual information. The process involves using the tokenizer on the textual information to construct a vocabulary and allocating distinct indices to individual words. Subsequently, the textual data is transformed into a series of numerical indices utilizing the acquired vocabulary. The process of tokenization holds significant importance in the preparation of textual data for the purpose of training deep learning models that demand numerical inputs.

The dataset is partitioned into training and testing sets using the '`train_test_split`' function from scikit-learn library. The testing set is allocated 20% of the data, while the training set is assigned the remaining 80%. Subsequently, the model is trained by utilizing the training data and corresponding labels. The '`validation_split`' parameter has been assigned a value of 0.2, indicating that a proportion of 20% of the training data will be allocated for validation purposes throughout the training iteration. This facilitates the monitoring of the model's efficacy on previously unseen data and serves to mitigate the risk of overfitting. The model will perform 8 iterations over the complete training dataset.

5.2 Logistic Regression

The utilization of logistic regression constitutes a fundamental aspect of our research. Despite logistic regression's limitations in capturing complex nonlinear relationships, it remains a valuable tool for detecting fake news. The popularity and practicality of this model stem from its interpretability, efficiency, probability estimation, and baseline performance. It is commonly used for initial explorations and serves as a benchmark for comparison with more advanced models. Logistic regression is a statistical method that yields results that are readily interpretable, thereby facilitating the comprehension and analysis of the variables that are influential in the identification of fake news. The model assigns coefficients to individual features, which serve as indicators of their respective strengths and directions of impact on the predicted outcome. The coefficients have the potential to be determined as a means of identifying the characteristics that have the greatest impact or words in discerning between authentic and fabricated news. This model offer probability estimates for individual instances, indicating the probability of their membership in a specific class (i.e., fake or real news). The process of probability estimation facilitates a more intricate comprehension of the level of assurance exhibited by the

model in relation to its prognostications. The model’s predictions can be fine-tuned to achieve a balance between precision and recall that aligns with the specific demands of the application by establishing a suitable threshold.

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}} \quad (5.1)$$

$P(y=1|x)$ in this equation denotes the likelihood that the dependent variable y will equal 1 given the input variables \mathbf{x} . The regression coefficients for the input variables $x_0, \beta_0, \beta_1, \beta_2, \dots, \beta_n$ are represented by the letters e , which stands for the base of the natural logarithm.

The linear combination of input variables and coefficients is transformed into a number between 0 and 1, which represents the expected probability of the positive class (in this example, $y=1$), using the sigmoid function $(1 / (1 + e-z))$. //

The logistic regression model proposes a linear association between the predictor variables and the logarithm of the odds of the response variable. Although this assumption may not be universally applicable in intricate situations, it can nevertheless prove efficacious in numerous instances. The identification of patterns and word usage that diverge from credible news sources is frequently employed in the detection of false news. In instances of this nature, a logistic regression model has the capability to proficiently acquire knowledge to distinguish between the two groups by utilizing a linear decision boundary. The use of logistic regression facilitates the identification of noteworthy features that play a role in the detection of fake news. Through an analysis of the size and trend of the coefficients, valuable insights can be obtained regarding the features or words that exert the greatest influence on the prediction. The aforementioned data has the potential to facilitate subsequent scrutiny and inquiry, thereby enabling professionals and researchers to gain a more comprehensive understanding of the attributes of fake news and possibly alleviate its ramifications.

The SMOTE (Synthetic Minority Over-sampling Technique) algorithm was employed in conjunction with logistic regression to address the issue of imbalanced datasets. The SMOTE class’s `fit_resample` method executes oversampling by producing synthetic samples from the minority class (fake news) to achieve class distribution equilibrium. The oversampled training data is stored in `X_train_smote` and `Y_train_smote`. Subsequently, the logistic regression model was trained using the resampled training data. `Y_pred` is the container for the predicted classifications (authentic or fabricated news) derived from the logistic regression model’s predictions.

5.3 Random Forest

The Random Forest technique is an ensemble learning approach that combines numerous decision trees to generate predictions. Through the process of integrating the forecasts of numerous trees, this approach capitalises on the collective intelligence of a group and mitigates the possibility of any single tree producing erroneous predictions. The utilisation of an ensemble approach serves to augment the comprehensive precision and resilience of the system designed for detecting fabricated

news. The Random Forest algorithm is designed to mitigate the problem of overfitting that may arise when using individual decision trees. This is accomplished through the incorporation of stochasticity during the training phase. Every individual tree within the forest is subjected to a training process that involves the selection of a random subset of the available training data and a random subset of features. The introduction of randomness serves to disassociate the individual trees and fosters heterogeneity among them, resulting in enhanced generalization and diminished overfitting.

$$\hat{y} = \text{RF}(x) = \frac{1}{N_{\text{trees}}} \sum_{i=1}^{N_{\text{trees}}} \text{tree}_i(x) \quad (5.2)$$

As determined by combining the predictions of many decision trees in the Random Forest, \hat{y} in this equation stands for the projected output. N_{trees} stands for the total number of trees in the forest, whereas x stands for the input characteristics. Each $\text{tree}_i(x)$ corresponds to the i -th decision tree's forecast.

The Random Forest algorithm is capable of effectively managing datasets that are of a large scale and contain a multitude of features and instances. Parallelization of the training process facilitates the usage of multi-core processors and distributed computing frameworks. The Random Forest algorithm's scalability and efficiency render it a viable option for handling substantial amounts of textual data that are frequently encountered in the context of fake news detection tasks.

The variable `X_train_2d` represents the input characteristics of the oversampled training dataset, `X_train_smote`, in a two-dimensional NumPy array. The function `np.stack` is utilised to concatenate the individual samples into a unified array vertically. The variable `X_train_2d` represents the test features organised as a two-dimensional NumPy array.

The training procedure involves using the oversampled training data to teach the random forest classifier, subsequently generating a prediction for the trained classifier.

5.4 KNN

The K-Nearest Neighbours (KNN) algorithm is a type of supervised learning approach that is applicable for classification machine learning tasks. The operational mechanism involves the allocation of a new data point to a category identity that exhibits the highest frequency among its proximate neighbours. The K-Nearest Neighbours (KNN) algorithm can be utilised for fake news detection by means of supervised learning on a labelled dataset. Every information point in the dataset represents a news article and its corresponding label, either real or fake. The algorithm acquires knowledge of the distinctive features and attributes of genuine and fabricated news articles, thereby facilitating its capacity to categorize novel and unobserved articles by their resemblance to the examples used for training. The K-Nearest Neighbours (KNN) algorithm is a robust and straightforward method that

can be utilized for analyzing both quantitative and qualitative data, rendering it a fitting choice for our research.

The K-Nearest Neighbours (KNN) algorithm is especially advantageous in the context of textual data analysis due to its reliance on the computation of similarity or distance metrics between individual data points. The present study involved the representation of every news article as a numerical feature vector through the utilization of methods including Word2Vec, FastText, or Gensim, as previously discussed. The K-Nearest Neighbours (KNN) algorithm utilises the feature vectors to compute the level of similarity among articles. Commonly, this task is accomplished by utilising distance metrics such as Euclidean distance or cosine similarity. The K-Nearest Neighbours (KNN) algorithm has the ability to generate predictions by analyzing the patterns present in the training data, specifically by examining the closest neighbours of a given article.

The K-Nearest Neighbours (KNN) algorithm yields results that are simple to comprehend due to its reliance on the majority vote of the closest neighbours to determine the predicted class. This feature facilitates the analysis of the designated adjacent articles of a particular news piece and comprehending the rationale behind its classification as authentic or counterfeit, contingent upon its resemblances with the data used for training. The power to interpret data can yield significant insights into the distinguishing characteristics and features of genuine and fabricated news in the Bangla language, thereby enhancing understanding of the issue at hand.

It is noteworthy that although KNN may serve as a proficient algorithm for detecting fake news, it entails specific considerations. The selection of the K parameter and distance metrics are crucial considerations that can significantly affect the performance of the K-Nearest Neighbour algorithm. Moreover, the management of data with a high number of dimensions or extensive datasets may present computational obstacles. Thus, in our study, we transformed the vector space with high dimensions into a two-dimensional vector space to alleviate the burden of handling high-dimensional data by KNN.

The classification of the news article is ascertained through a process of majority voting among the K nearest neighbours. The classification of a news article as either real or fake is determined by the labelling of the majority of its K nearest neighbours. Specifically, if the majority of these neighbours are labelled as real, then the news article is classified as real; otherwise, it is classified as fake. Here is the equation for the KNN,

$$\hat{y} = \text{KNN}(x) = \text{mode}(\{y_i\}_{i=1}^k) \quad (5.3)$$

\hat{y} in this equation stands for the expected result given an input x . The input sample is given a class label by the KNN algorithm based on the k nearest neighbours most prevalent class label (mode).

In order to discover the k nearest neighbours to the input sample, the KNN method uses a distance metric (like Euclidean distance). The most common class label is chosen as the projected class for the input after looking at the class labels of these

k neighbours.

5.5 Support Vector Machine

The classification of fake and real news can be achieved through the use of Support Vector Machine (SVM). This method involves the identification of an optimal hyperplane that effectively separates the two classes. The features extracted from the news articles are used to inform this process. The Support Vector Machine (SVM) algorithm is a robust method capable of efficiently managing high-dimensional data and identifying the optimal hyperplane for segregating distinct classes. Furthermore, Support Vector Machines (SVM) exhibit high efficacy in scenarios where the data points demonstrate linear separability, indicating the presence of a distinct boundary between distinct classes. Within the realm of fake news identification, Support Vector Machines (SVM) possess the capability to acquire knowledge of a decision boundary that distinguishes genuine news articles from fabricated ones, contingent upon the features extracted from the data. The support vector machine (SVM) algorithm seeks to identify an optimal hyperplane that maximises the margin between classes. This approach has the potential to enhance generalisation and classification performance.

The SVM model is trained using the feature vectors of the labelled training examples. The objective of Support Vector Machines (SVM) is to identify the optimal hyperplane that maximises the difference between the classes, while minimising the number of incorrect classification errors. The decision boundary region that distinguishes between fake and real news articles is represented by the hyperplane in SVM. The objective of Support Vector Machines (SVM) is to identify the hyperplane that optimises the margin, defined as the separation distance between the hyperplane and the closest training instances from every class. The positioning of the hyperplane is optimised to achieve highest separation between the two distinct classes, while simultaneously preserving a maximum margin. The utilisation of a suitable kernel function in SVM enables the capture of intricate relationships and identification of non-linear decision boundaries, thereby enhancing its capacity to differentiate between genuine and fabricated news articles. The SVM algorithm is designed to prioritise the support vectors over all other training examples. This approach enhances the algorithm's memory efficiency and effectiveness, particularly when dealing with high-dimensional data.

Upon completion of the training process, the Support Vector Machine (SVM) model is capable of categorizing novel and unobserved news articles. The feature vectors of the newly generated articles were subjected to the same transformation techniques that were utilised during the training phase. The Support Vector Machine (SVM) algorithm is utilized to classify a new article by determining its position relative to the hyperplane.

The classification of a news article as either fake or real is determined by SVM based on its placement relative to the decision boundary.

5.6 Gaussian Naive Bayes

The Gaussian Naive Bayes algorithm is a type of Naive Bayes technique that is based on the assumption of the Gaussian distribution for the continuous features present in the data. The Gaussian Naive Bayes algorithm is commonly used in research for datasets that consist of a variety of numerical features. These features may include TF-IDF scores, word embeddings, or other continuous ways to represent text data. The Word2Vec models were utilised in the course of our research. The research anticipates that the feature values for every category follow a normal distribution. The Gaussian Naive Bayes algorithm employs statistical estimation techniques to determine the mean as well as the standard deviation of every feature for both the fake and real classes, utilising the training data.

The research assumes that the feature values for each class adhere to a Gaussian (normal) distribution. The probability density function (PDF) is computed for each feature and class by utilising the estimated median and standard deviation. The probability density function (PDF) is a statistical representation of the probability of observing a specific feature value, provided a specific class of features.

$$P(C|x) = \frac{1}{Z}P(C) \prod_{i=1}^n P(x_i|C)$$

In the classification phase, the Gaussian Naive Bayes algorithm employs Bayes' theorem to compute the posterior probability of each class based on the identified features. According to Bayes' theorem, the posterior probability of class C given the observed features X can be calculated using the formula $P(C|X) = (P(X|C) * P(C))/P(X)$. This formula involves the prior probability of class C, denoted as P(C), and the likelihood of observing the features X given the class C, denoted as P(X | C). The denominator, P(X), represents the probability of observing the features X.

The classification of a new news article using Gaussian Naive Bayes is based on the posterior probability of each class. This probability is calculated by considering the observed features and determining the likelihood of the article belonging to each class. The class regarding the greatest posterior probability is then assigned to the article as the most probable class. The news article is classified as either fake or real through a decision line that is determined by the posterior probability. The determination of the decision boundary is based on the analysis of the mean and standard deviation for the features for every class.

5.7 Bagging

The classification technique known as Bagging (Bootstrap Aggregating) has been employed to classify news articles as either real or fake, utilising their respective features. Bagging is a widely used ensemble learning technique in which a set of classifiers are trained on various subsets of the data, and their predictions are combined to produce a final output. This method has been shown to improve the accuracy and robustness of machine learning models, and is commonly used in a variety of applications.

The bagging technique is a resampling method that generates several bootstrap samples by randomly selecting subsets from the original training dataset. The subsets were generated through a process of random sampling with substitutes compared to the primary dataset. The amount of data of each and every portion in the dataset is subject to variation, although it is typically equivalent to the original dataset.

In the classification phase, a set of trained classifiers are utilized to classify newly acquired news articles. The predictions generated by the classifiers are based solely on the observed characteristics of the news article. In the context of binary classification, specifically in the domain of distinguishing between fake and real news, each classifier operates autonomously to predict the respective class. The final prediction is made by aggregating the predictions generated by the individual classifiers. The final prediction is determined by selecting the class with the largest aggregated probability or vote by a majority.

5.8 Boosting

The ensemble learning method known as Boosting has been investigated as a potential means of classifying news articles as either real or fake, utilising their respective features. Boosting is a machine learning technique that differs from Bagging in that it trains insufficient classifiers in a sequential manner to generate an effective classifier. In contrast to Bagging, that trains several classifiers independently, Boosting iteratively improves the performance of the weak classifiers by focusing on the misclassified instances in each iteration. This process continues until the desired level of accuracy is achieved.

The boosting technique involves the iterative training of a set of weak classifiers, which are usually decision trees. In each iteration, the weak classifier is trained with the objective of minimising the weighted deviation. The weights assigned to the incorrectly classified instances from the previous iterations of training are given more emphasis. The iterative process is executed for a predetermined number of iterations after which a specific performance threshold is attained. Instances that are misclassified are given greater weights to enhance their significance in the following iteration. In order to improve the accuracy of classification, it is generally accepted to reduce the weights of correctly classified instances. This approach allows for a greater focus on challenging instances, which may be more difficult to classify accurately

In the classification phase, the final prediction is made by combining the predictions of each weak classifier. The conventional approach for making the final prediction involves utilising a weighted majority voting or weighted averaging mechanism. The weights assigned to the weak classifiers are proportional to their accuracy, with more accurate classifiers receiving higher weights.

Chapter 6

Result Analysis

Assessment of the performance and effectiveness of classification models necessitates the use of evaluation metrics as indispensable instruments. Quantitative measures are utilized to aid researchers and practitioners in comprehending the efficacy of a model in accurately categorizing instances into distinct classes, such as distinguishing between fabricated news and authentic news.

6.1 Accuracy, precision, recall, and F1:

Accuracy, precision, recall, and F1 score are widely utilized evaluation metrics that are essential for assessing the overall effectiveness of classification models. The metric of accuracy evaluates the ratio of accurately classified instances to the overall instances present in the dataset.

The metric offers a comprehensive assessment of the accuracy of the model's predictions for both positive and negative cases. The concept of precision pertains to the affirmative category (specifically, fabricated news) and gauges the ratio of accurate positive prognostications (i.e., appropriately identified counterfeit news) to the overall anticipated positive occurrences. The degree of precision is indicative of the model's capacity to minimize the occurrence of false positives.

Models	Precision	Recall	F1 Score	Accuracy
LSTM	0.91	0.91	0.91	0.91
Logistic Regression	0.87	0.89	0.88	0.90
Support Vector Machine	0.88	0.89	0.88	0.90
Random Forest	0.88	0.89	0.89	0.91
Gaussian Naive Bayes	0.77	0.81	0.89	0.81
K-Nearest Neighbors	0.83	0.88	0.85	0.90
Bagging	0.86	0.88	0.87	0.89
Boosting	0.83	0.85	0.83	0.86

Figure 6.1: Scores in Tabular format

The tabulated data indicates that the LSTM model outperforms other models with an accuracy score of 0.91. Moreover, the Logistic Regression, Support Vector Machine, and Random Forest classifiers exhibit favorable accuracy, precision, and recall metrics subsequent to the implementation of LSTM. Although the other two models exhibit comparable accuracy, Random Forest outperforms them by a slight margin with an accuracy score of 0.91.

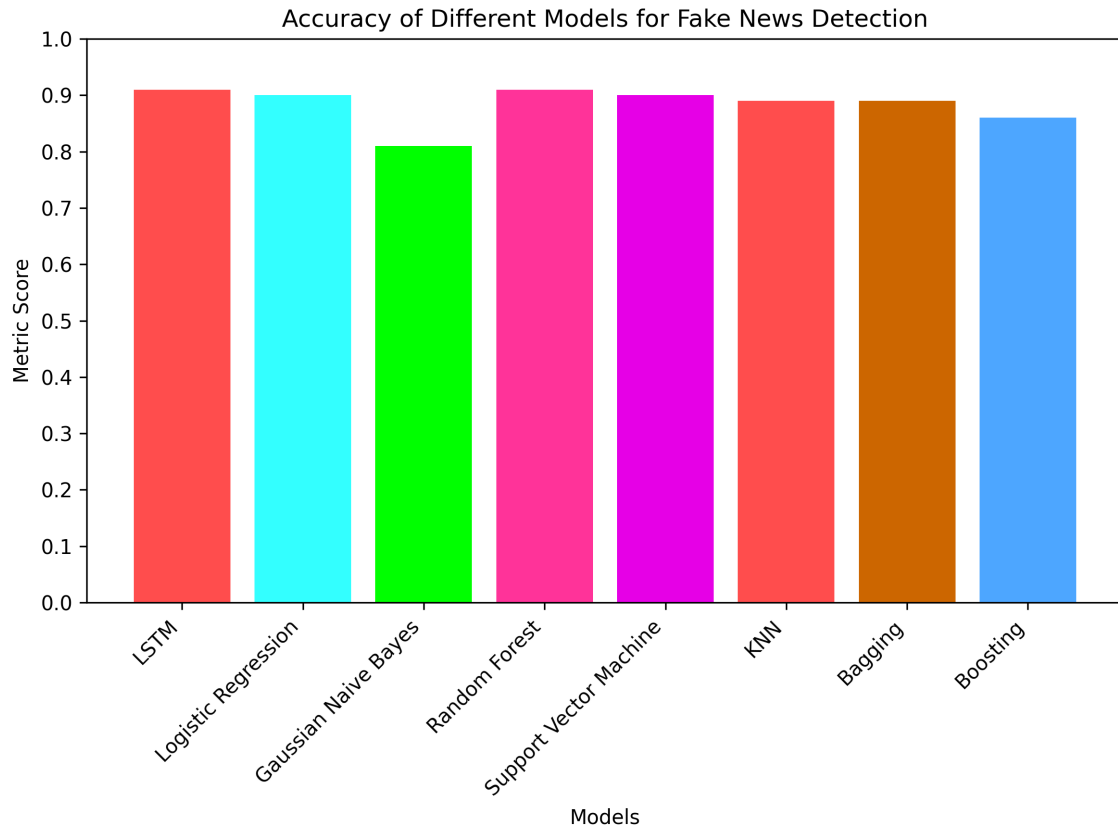
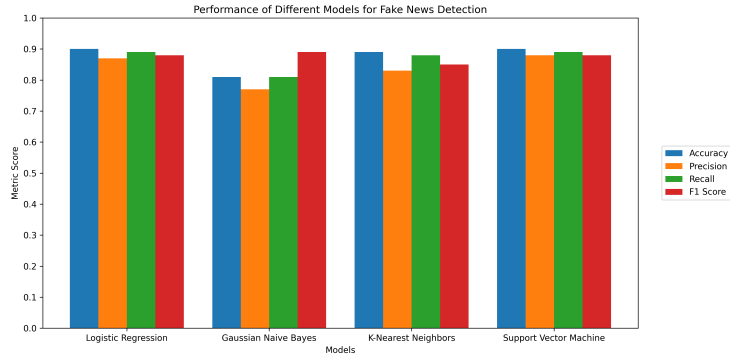


Figure 6.2: Scores in Tabular format

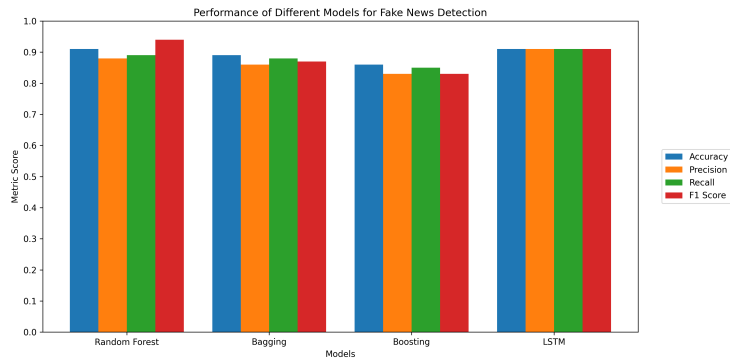
The models exhibited comparable levels of accuracy, suggesting uniform performance across various methodologies. The precision scores exhibited considerable variation among the models, thereby underscoring the dissimilarities in their capacity to accurately detect affirmative cases (i.e., fake news). Several models demonstrated varying levels of proficiency in recall, effectively capturing a greater percentage of authentic positive instances (i.e. fake news), whereas others encountered difficulties in achieving the same outcome, including Random Forest, LSTM, and KNN.

The importance of mitigating misinformation necessitates the utilization of models with elevated precision, such as LSTM, SVM, Random Forest, and Logistic Regression, to accurately detect fabricated news and minimize erroneous identifications. The enhanced recall attained by particular models indicates their efficacy in apprehending a larger quantity of fake news instances, which is crucial in defending against misinformation.

The findings of our study indicate that the LSTM model outperforms other models by a significant margin, while the Gaussian Naive Bayes model exhibits comparatively inferior performance. The Long Short-Term Memory (LSTM) is a recurrent neural network architecture that is adept at modeling sequential data such as text.



(a) Score chart 1



(b) Score chart 2

Figure 6.3: Accuracy, precision, recall, and F1 score chart

This makes it a highly suitable tool for tasks such as detecting fake news. Notwithstanding their intricate architecture, LSTMs can pose a computational burden, necessitating substantial computational resources and training duration. Conversely, the Gaussian Naive Bayes algorithm is a less complex probabilistic classifier that is predicated on a robust assumption of feature autonomy. The method exhibits computational efficiency and demonstrates reduced training and inference times. Hence, in the context of our study, Gaussian Naive Bayes could potentially outperform LSTM in terms of computational efficiency.

LSTM models are renowned for their adeptness in capturing intricate patterns and dependencies in sequential data, thereby resulting in commendable generalization performance. Consequently, it is plausible that LSTM models exhibit a greater probability of achieving favorable results when presented with novel and unfamiliar instances of fabricated news. Conversely, the Gaussian Naive Bayes algorithm is predicated on the presumption of feature independence, potentially constraining its capacity to apprehend intricate associations within the data. As a result, the model may encounter difficulties in extending its performance to novel cases that differ from the assumption of independence. In general, it is widely acknowledged that LSTM models exhibit superior generalizability in comparison to Gaussian Naive Bayes.

6.2 Confusion Matrix:

The employment of the confusion matrix is a pivotal technique in assessing the efficacy of classification models. The output presents a tabular format that illustrates the model's predicted values in comparison to the factual class labels of the given dataset. The matrix provides a comprehensive overview of the results obtained from the classification procedure, facilitating an in-depth evaluation of the precision of the model and the types of errors committed.

It generally comprises of four values: The True Positive (TP) refers to the count of accurately predicted positive instances, specifically in the context of identifying fake news. The True Negative (TN) refers to the accurate identification of negative instances, specifically the correct prediction of real news. The False Positive (FP) refers to the count of positive instances that have been inaccurately predicted (misclassified as fake news despite being genuine news). The False Negative (FN) refers to the count of negative instances that have been inaccurately predicted (misclassified as genuine news despite being fake news).

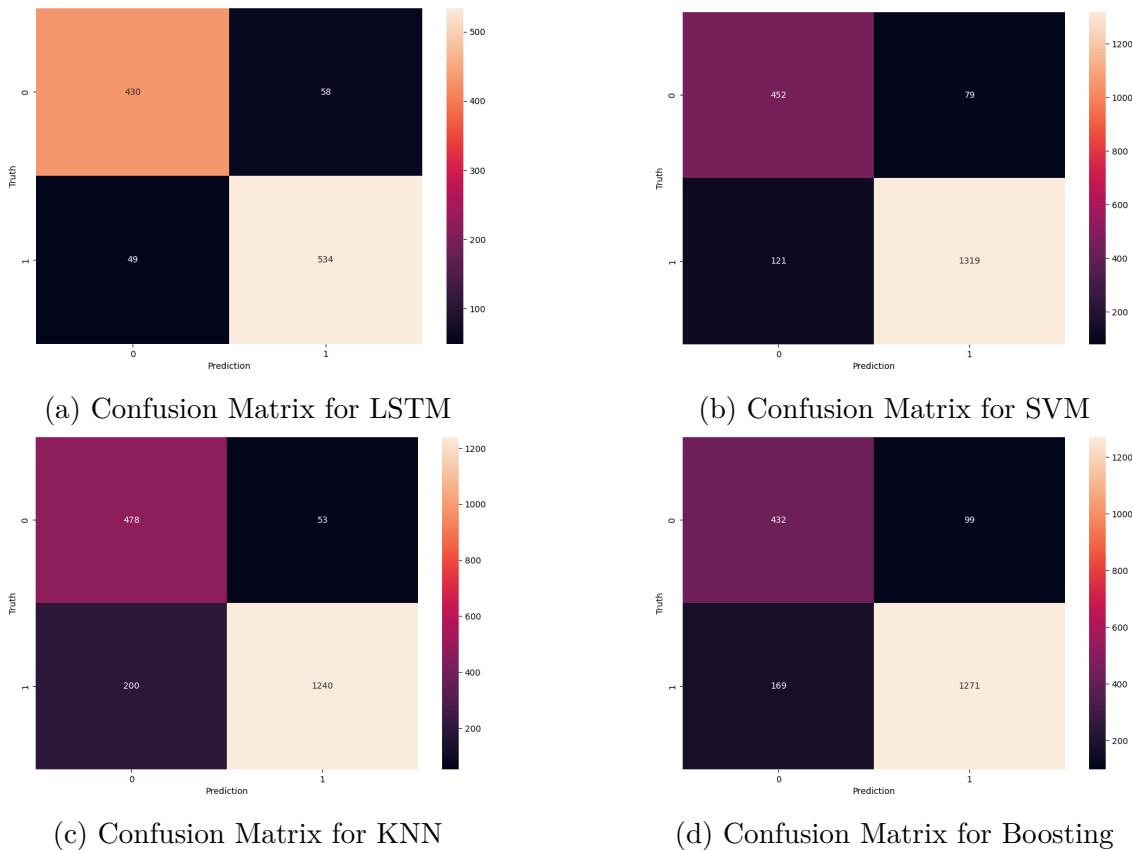


Figure 6.4: Confusion Matrices

The LSTM model utilized in our research accurately classified 430 instances of fabricated news, while erroneously misclassifying 58 instances of fabricated news as genuine news. Furthermore, the aforementioned model accurately categorizes a total of 534 news articles as genuine, while erroneously labeling 49 authentic news pieces as counterfeit. In comparison to the Support Vector Machine, this model exhibits a lower level of performance, as evidenced by its misclassification of 79 instances of fake news as real news and 121 instances of real news as fake news. Thus,

it can be inferred that the LSTM model exhibits a marginally superior performance compared to the Support Vector Machine.

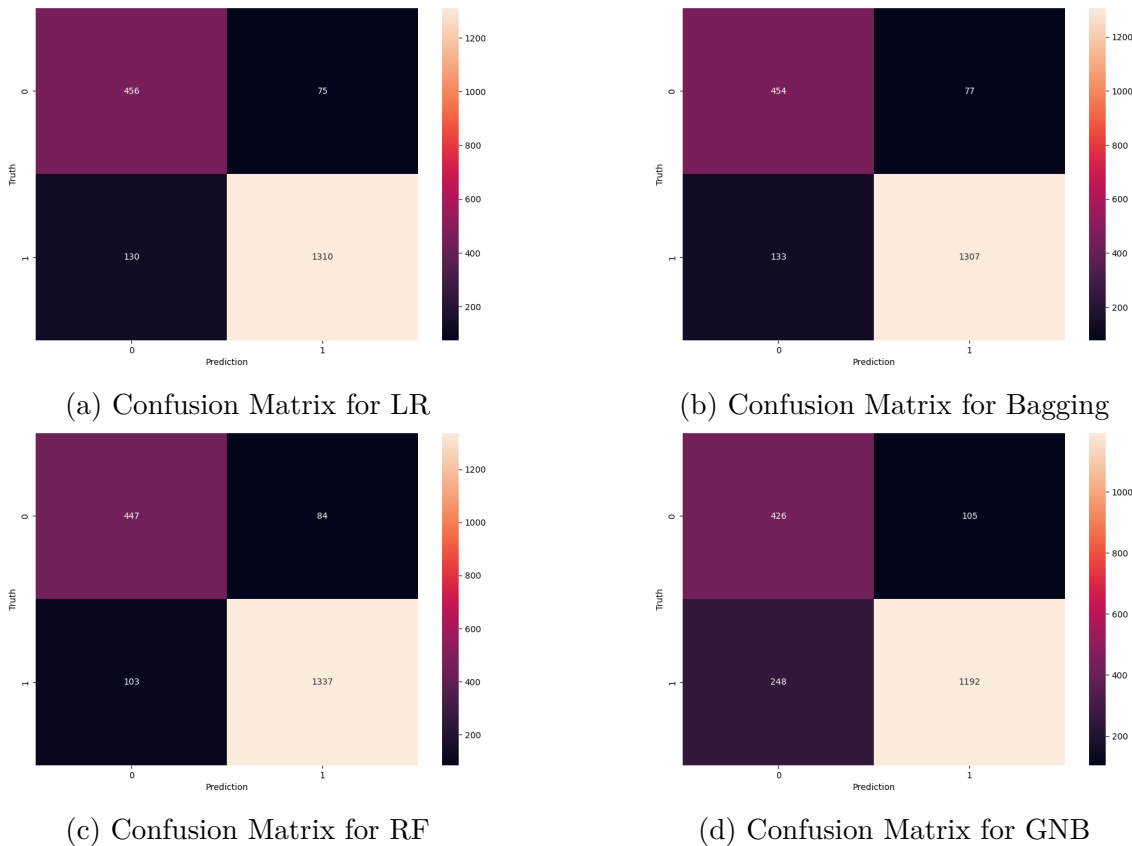


Figure 6.5: Confusion Matrices

The confusion matrix of Random Forest indicates a low rate of misclassification, with only 84 instances of fake news and 103 instances of real news being misclassified. In contrast, a high rate of correct classification was observed, with 447 instances of fake news and 1337 instances of real news being classified accurately. Conversely, Logistic Regression exhibited a low error rate in misidentifying a mere 75 instances of fake news and 130 instances of real news. In contrast, accurately categorizing 456 instances of fabricated news and 1310 instances of authentic news. Hence, discerning the efficacy of Logistic Regression and Random Forest algorithms based on the confusion matrix poses a significant challenge. While the Random Forest algorithm exhibits a slightly superior performance compared to Logistic Regression.

6.3 ROC Curve:

The Receiver Operating Characteristic (ROC) curve is a frequently employed assessment metric in the field of machine learning, encompassing artificial intelligence models. Binary classification problems, such as the detection of fake news, can be effectively addressed through the use of this method. The primary objective is to accurately classify instances into one of two distinct categories: either fake or genuine news.

The Receiver Operating Characteristic (ROC) curve is a graphical representation of a classifier's effectiveness, which displays the sensitivity or recall (true positive

rate) against the specificity (false positive rate). The true positive rate pertains to the ratio of accurately classified positive instances, specifically those that pertain to fake news, in relation to the total number of positive instances. Conversely, the false positive rate pertains to the ratio of inaccurately classified negative instances, specifically those that pertain to genuine news, in relation to the total number of negative instances.

The process of constructing a Receiver Operating Characteristic (ROC) curve involves utilizing the predictions of a classifier for various threshold values. By manipulating the threshold value that determines the classification of a prediction as positive or negative, it is possible to derive distinct true positive rates and false positive rates. The Receiver Operating Characteristic (ROC) curve illustrates the balance between the true positive rate (sensitivity) and the true negative rate (specificity) in relation to the variation of the decision threshold.

Each point on the Receiver Operating Characteristic (ROC) curve corresponds to a specific threshold value. The curve is derived by linking these data points. An optimal classifier would exhibit a sensitivity of 100% (true positive rate of 1) and a specificity of 0% (false positive rate of 0), thereby yielding a coordinate at the uppermost left-hand corner of the Receiver Operating Characteristic (ROC) space. In contrast, a random classifier would generate a diagonal line extending from the lower left to the upper right quadrant.

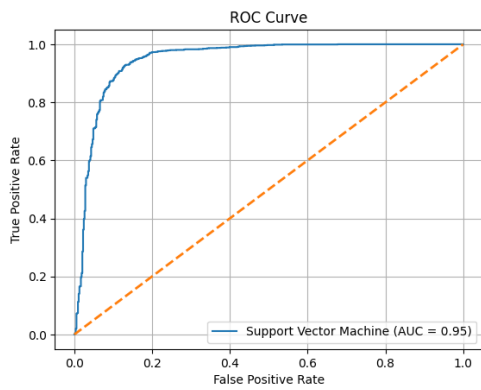
The utilization of the ROC curve's area (AUC-ROC) is a prevalent statistical measure that evaluates the comprehensive efficacy of a classifier. An ideal classifier would exhibit an AUC-ROC value of 1, signifying its ability to accurately differentiate between the two classes without any errors. A classifier that operates randomly would exhibit an AUC-ROC value of 0.5, as its efficacy would be similar to random guessing.

The Receiver Operating Characteristic (ROC) curve facilitates the graphical evaluation of numerous classifiers or models. Models exhibiting a curve in proximity to the upper left corner or a greater AUC-ROC metric are deemed superior. All the model curves are presented herein for the reader's evaluation. The model that exhibits the highest AUC-ROC value is the Support Vector Machine (SVM) and LSTM, with a score of 0.95.

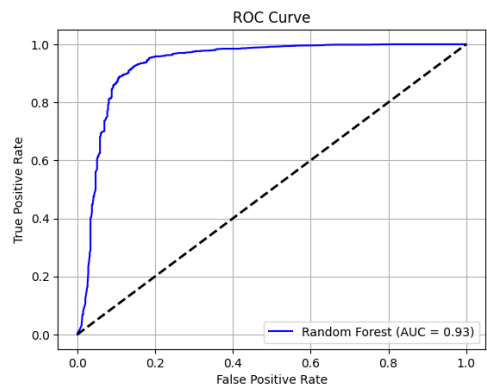
The Receiver Operating Characteristic (ROC) curve is helpful in determining the optimal threshold by considering the balance between sensitivity and specificity. As an illustration, it can be observed that K-Nearest Neighbors (KNN) and Random Forest classifiers exhibit identical Area Under the Curve (AUC) scores. However, it can be argued that KNN is expected to outperform Random Forest due to its relatively superior ability to optimize true positive outcomes. Bagging is seen to have an advantage in this case over Boosting in reducing false positives based on the threshold. The curves exhibit resilience towards class imbalance, thereby enabling it to proficiently assess models even in scenarios where the number of cases in every class is distributed differently.

Through the examination of the ROC curve's configuration and dynamics, one can acquire valuable knowledge regarding the model's performance attributes, including its capacity to manage diverse forms of inaccuracies and its comprehensive discriminatory efficacy.

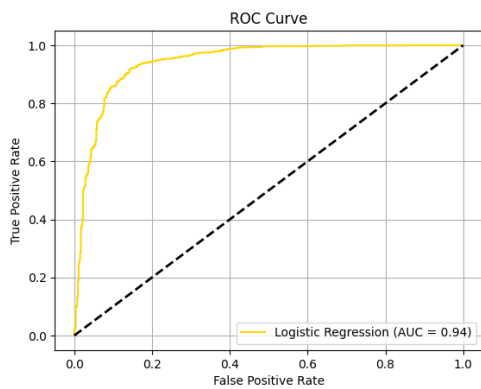
Although the ROC curve and AUC-ROC offer significant insights into the perfor-



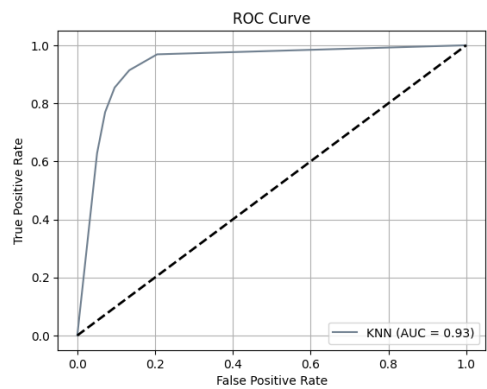
(a) ROC for Support Vector Machine



(b) ROC for Random Forest

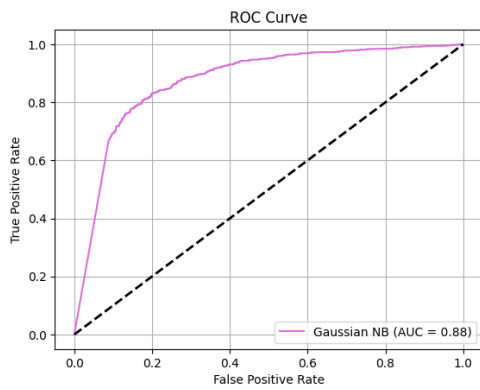


(c) ROC for Logistic Regression

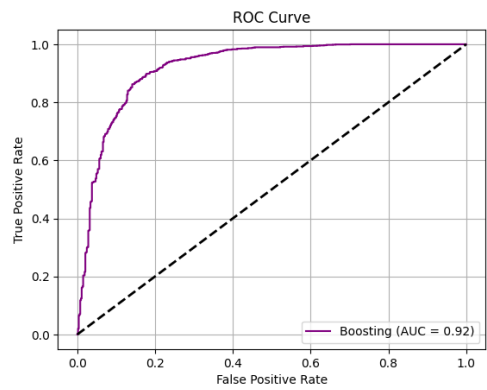


(d) ROC for K-Nearest Neighbor

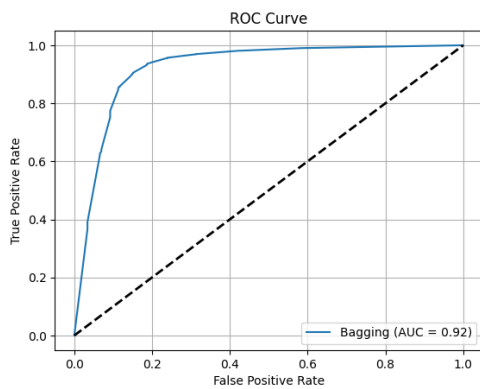
Figure 6.6: Receiver operating characteristic Curves



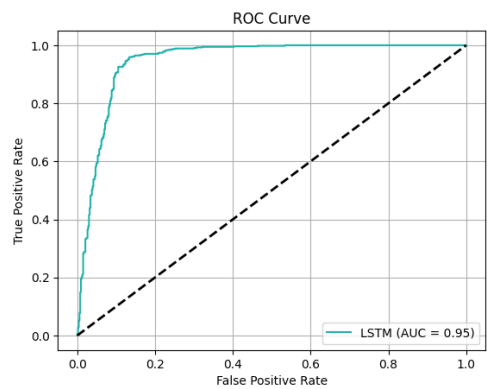
(a) ROC for Gaussian Naive Bayes



(b) ROC for Boosting



(c) ROC for Bagging



(d) ROC for LSTM(Deep Learning)

Figure 6.7: Receiver operating characteristic Curves

mance of classifiers, it is crucial to acknowledge that they concentrate on the general classification performance and do not encompass data regarding the particular decision threshold or the expenses related to misclassifications. Hence, it is imperative to take into account additional evaluation metrics and domain-specific prerequisites while evaluating the appropriateness of an artificial intelligence model for a specific application.

6.4 Learning Curve:

The learning curve is a visual depiction of the proficiency of an artificial intelligence model in relation to the magnitude of the training data utilized. Gaining insight into the manner in which the performance of the model progresses with an increase in the amount of data utilized for training can prove to be beneficial. The utilization of learning curves is a prevalent practice in machine learning for the purpose of evaluating model performance, detecting potential problems such as underfitting or overfitting, and employing evidence-based strategies for training models and data requirements.

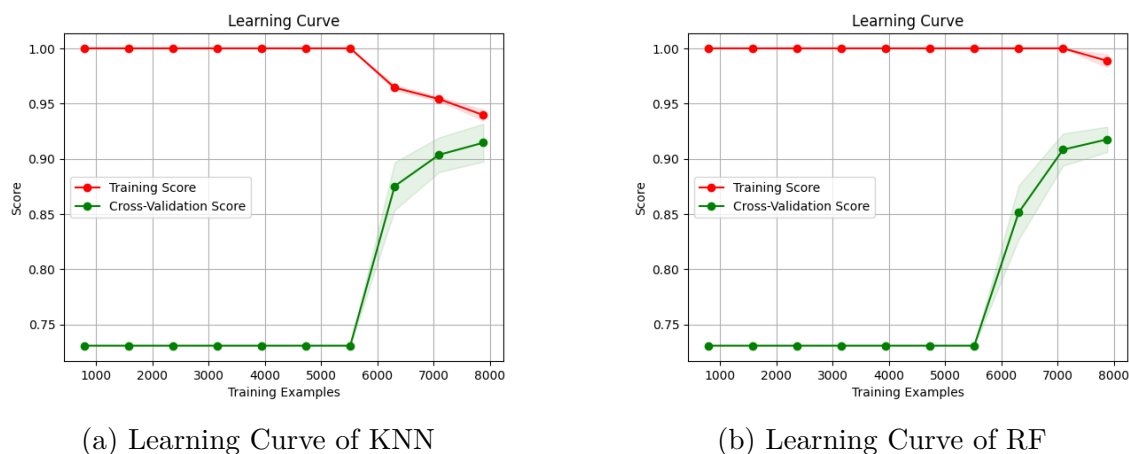
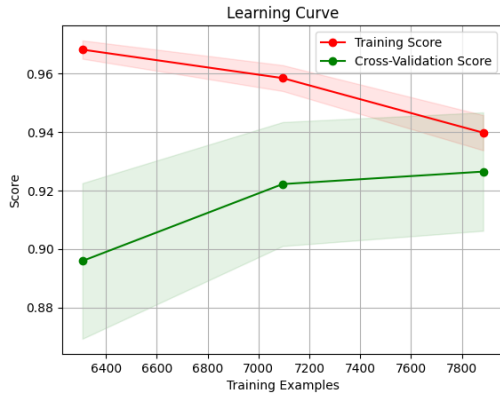
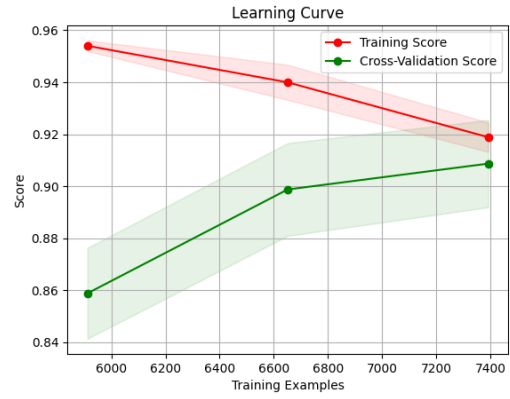


Figure 6.8: Learning Curves comparison 1

When the performance of the training set is significantly greater than that of the validation set, it suggests the presence of high variance or overfitting. Conversely, in the event that the performance of both the training and validation phases is suboptimal, it indicates a propensity towards high bias or underfitting. Upon comparing the performance of K-Nearest Neighbors (KNN) and Random Forest algorithms, it can be inferred that KNN exhibits a superior convergence rate in comparison to Random Forest. Therefore, the predictions generated by KNN are deemed more reliable than those generated by Random Forest. The Random Forest model exhibits a high degree of variance, indicating a tendency towards overfitting.



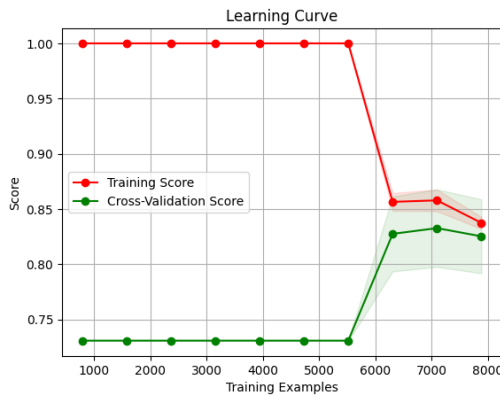
(a) Learning Curve of SVM



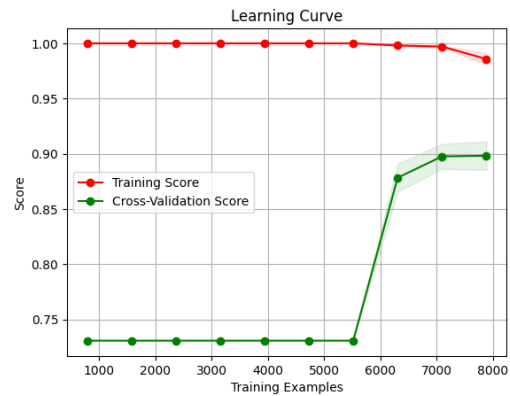
(b) Learning Curve of LR

Figure 6.9: Learning Curves comparison 2

The analysis of learning curves is crucial in assessing the adequacy of the training data at hand. When the learning curve indicates that the model's performance has reached a plateau and further data does not result in significant enhancements, it implies that the model has attained its maximum learning potential from the existing data. The present analysis has the potential to provide guidance in making decisions pertaining to the acquisition of additional training samples or evaluating the necessity of more intricate models. The SVM and Logistic Regression models exhibit favorable convergence rates. However, it is noteworthy that the SVM model has reached a plateau, suggesting that an adequate amount of data has been fed into the model. The models exhibit a favorable balance between variance and bias, resulting in a reliable prediction score.



(a) Learning Curve of GNB



(b) Learning Curve of Bagging

Figure 6.10: Learning Curves comparison 3

The utilization of cross-validation warrants a learning curve that exhibits uniform proficiency over numerous folds. This suggests that the efficacy of the model is not excessively influenced by the particular partitioning of the data and exhibits resilience across various subgroups of the dataset. The graphical representation of the Gaussian Naive Bayes and Bagging classifiers indicates that these models are not suitable for our classification task. The utilization of bagging learning rate can potentially lead to overfitting, which is a common issue in machine learning. Despite exhibiting a favorable convergence rate, the Gaussian Naive Bayes classifier

lacks an optimal equilibrium between variance and bias. The model's performance is expected to be suboptimal as a result of reduced accuracy.

Drawing upon both visual and quantitative analyses, it can be determined that SVM, Logistic Regression, and KNN exhibit greater potential than the other machine learning models utilized in this study. Nonetheless, there exists a potential for enhancement.

Chapter 7

Conclusion

7.1 Challenges

The pre-training of large language models demands a substantial amount of superior text data. As an example, BERT [11] has undergone pre-training on the English Wikipedia and the Books corpus, which comprises 3.3 billion tokens. The Bangla language is characterized by limited resources, as evidenced by the comparatively small size of the Bangla Wikipedia dump from July 2021, which is only 650 MB. This is notably smaller than the English Wikipedia, with a difference of two orders of magnitude. Consequently, there exists an insufficiency of data to facilitate the training of a sizable model and achieve a precise outcome. The Bangla language is comparatively less developed in the field of Natural Language Processing research when compared to English. Considerable research has been conducted in the field of Bangla Language processing. For instance, Hossain’s BanFakeNews [17] project involved an analysis of the dataset and the development of a benchmark system utilizing advanced NLP techniques to detect false news in Bangla. BanglaBERT is a Natural Language Understanding (NLU) model that is based on BERT and has been pre-trained in Bangla, a language that is frequently spoken and well-known in the NLP domain, despite being limited in resources. The purpose of this investigation is to advance the field of Fake News Detection through the utilization of the aforementioned two studies. Nonetheless, there is a scarcity of extensive libraries and pre-trained models such as BNLP and banglaBERT.

The development of a proficient machine learning model calls for a significant quantity of precisely categorized training data. The lack of adequate Bangla data for detecting fake news poses a significant challenge in training a resilient model, resulting in difficulties in this field. Insufficient data can lead to overfitting or inadequate depiction of the various patterns that exist in fabricated news.

The development of Natural Language Processing (NLP) techniques for Bangla can present challenges due to the limited availability of linguistic resources and tools. The accessibility of superior corpora, lexicons, and pre-existing models tailored towards Bangla are restricted. The limited availability of data hinders the advancement of more advanced methodologies and approaches for detecting fabricated news in the Bengali language.

The language of Bangla exhibits a high degree of complexity, characterized by a multitude of dialects and regional variations. The phenomenon of fake news may manifest itself through intricate linguistic nuances and subtleties, which may necessitate a profound comprehension of the language in question. Accurately addressing these intricacies within the model poses a greater challenge in cases where there are limitations in linguistic resources and language-specific NLP techniques.

The efficacy of fake news detection models is enhanced by the availability of domain-specific data that encompasses a wide range of news topics and types. The limited diversity of available data may constrain the model’s capacity to generalize to various categories of fabricated news. Compiling an all-encompassing and inclusive dataset that encompasses a diverse array of subjects in the Bangla language can pose a significant challenge.

The realm of fabricated news is in a state of perpetual flux, characterized by the frequent emergence of novel methodologies and strategies. In order to ensure precise detection, it is imperative that the model undergoes periodic updates and evaluations utilizing current data. In situations where resources and attention are limited, maintaining the model’s currency and efficacy in detecting contemporary instances of fabricated news can prove to be a formidable task.

7.2 Limitations

The incorporation of BLTK, a toolkit for Natural Language Processing in Bengali [26], has demonstrated its utility in the processing of specific portions of our dataset. The toolkit offers a variety of features that are tailored to facilitate the analysis of Bengali text. Nevertheless, it has been noted that the implementation and application of this library face intermittent challenges within our instructional setting. The primary reason for this is the inactive status of the BLTK library, which receives sporadic maintenance and updates, leading to compatibility challenges with modern iterations of the scikit-learn library. The irregularity of updates impedes the seamless operation of BLTK and presents obstacles when integrating it into our research workflow.

Furthermore, in the preprocessing stage, we faced difficulties while implementing stemming methodologies on our training dataset [18]. The dearth of suitable stemmers for Bengali text processing has resulted in inadequate resources and suboptimal performance with regards to the accuracy of stemming. The stemmers utilised in our dataset on occasion result in the loss of noteworthy information and semantic comprehension of the data. The restricted nature of stemming techniques poses a challenge to their efficacy in comprehensively capturing the complete essence of Bengali text, which could potentially impede the overall performance of our models.

Moreover, there is a scarcity of models accessible for the vectorization of Bengali textual material. Although attempts were made to employ the available alternatives, the constrained assortment curtails our capacity to investigate and broaden our strategy towards feature representation. The availability of a wider range of models that are customised for Bengali text vectorization would facilitate the ex-

ploration of diverse techniques and potentially augment the efficacy of our models.

Ultimately, a crucial matter arose in the conservation of our deep learning model that was constructed utilising the Keras framework. Although the model’s performance is satisfactory in its native environment, we faced challenges when endeavouring to preserve its parameters during local storage. The aforementioned matter could potentially be ascribed to discrepancies in compatibility or versioning between the Keras library and the configuration of the host system. It is expected that the matter at hand will be resolved in upcoming updates to the library, as the Keras community is currently engaged in addressing compatibility issues and enhancing the overall user experience.

Ultimately, a pivotal matter was confronted in the conservation of our deep neural network model, which was constructed utilising the Keras framework. Although the model exhibits satisfactory performance within its native setting, we faced challenges when endeavouring to preserve its parameters during local storage. The aforementioned matter could potentially be ascribed to discrepancies in compatibility or versioning between the Keras library and the configuration of the system in question. It is expected that forthcoming updates to the library will address this issue, as the Keras community is actively engaged in resolving compatibility issues and enhancing the overall user experience.

The aforementioned challenges and limitations underscore the necessity for continuous advancement and enhancement in the domain of Bengali natural language processing. The resolution of these concerns would significantly augment the dependability and efficacy of our research, empowering us to more effectively utilise the existing tools and methodologies for scrutinising Bengali text and propelling the progress of this field.

7.3 Discussion and Future Work

Despite the significant progress made in detecting fake news in the English language over the past decade, research on detecting fake news in the Bangla language remains limited. Our research aims to make a contribution to the Bangla Language by developing a method for detecting fake news articles.

Following the completion of our machine learning and deep learning models, we proceeded to create a website that is accessible to users and presents insights derived from our research. The objective of the website is to facilitate users in swiftly and conveniently ascertaining the genuineness of a Bangla news article. Utilising our top-performing model, which has exhibited exceptional precision in differentiating between fabricated and authentic news, individuals can acquire valuable insights into the dependability of the data they come across.

The website’s usability is characterised by its straightforward and intuitive design. To obtain Bangla news articles, users can utilise our platform by copying and pasting articles from reputable sources into the search box. The model employs its acquired knowledge and predictive abilities to examine the material and produce an outcome

that determines whether the news is categorised as genuine or fabricated. The ability to make informed decisions about the credibility and trustworthiness of news is crucial for users.

During the research phase, the model's performance was thoroughly tested and evaluated to determine its effectiveness in accurately classifying news articles. The model's robustness and reliability were ensured through careful training on diverse and representative datasets, incorporation of relevant features, and utilisation of state-of-the-art techniques in natural language processing and machine learning.

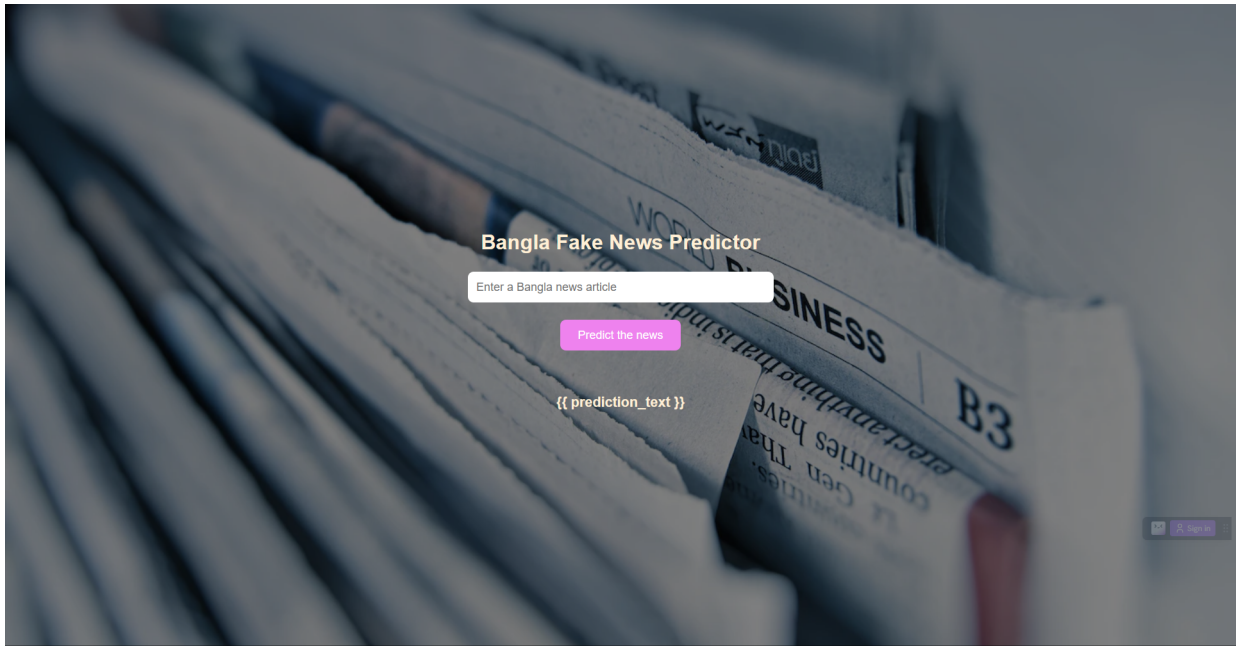


Figure 7.1: Implementation of our model to a website

The ongoing development of our implementation is aimed at enhancing the reliability and suitability of our research for public use. We are dedicated to ensuring that our research remains continuously improved to meet the highest standards of quality. The research is centred not only on evaluating the performance of machine learning and deep learning models, but also on analysing the underlying components that play a role in their efficacy.

Comprehensive assessments and evaluations of various Bangla word vectorizer models were conducted in the pursuit of optimal performance. The study conducted an evaluation of various models and found that the word2vec model developed by Sagor Sarkar[6] exhibited exceptional performance, outperforming other available Bangla word vectorizers. The results of this study underscore the possibility of utilising cutting-edge methodologies and specialised models in order to improve the precision and resilience of our models.

Although, the dedication to progress implies that we are not satisfied with relying solely on this achievement. The team is currently engaged in active research and experimentation with vectorizer models that are tailored to the unique characteris-

tics of the Bangla language. The objective of this study is to exploit the advantages of various models, recognise probable constraints, and explore innovative techniques that can advance the dependability and accuracy of machine learning and deep learning models.

Our research is focused on establishing a strong basis of dependability and efficiency, which is being achieved through continuous endeavours. The study aims to provide practical outcomes and understandings that can be implemented in practical situations, enabling individuals to make informed judgements and counteract the proliferation of false information in the Bengali language. The research team is dedicated to advancing the field of machine learning applications in Bangla language processing through a commitment to continuous improvement. Their work contributes to the broader knowledge and understanding of this field.

Furthermore, the present study emphasises the importance of enhancing the functionality of our website to offer users a more comprehensive and dependable evaluation of news articles, in addition to its existing focus on predictive capabilities of our model. In order to enhance our predictive capabilities, our research aims to integrate supplementary functionalities and refining mechanisms that surpass the prognostication derived solely from our machine learning and deep learning algorithms.

The proposed enhancement involves the addition of a manually operated checker that incorporates various tiers of filters. The study aims to assess the validity of news articles by conducting a verification process through trustworthy and reputable sources such as Prothom Alo, Jago News, BDNews24, and other established news outlets. Through cross-referencing the content with established sources, our website aims to enhance the validation process and bolster the credibility of the classification process. The website's classification of a news article as authentic is contingent upon its presence in reputable sources, according to research. This approach is intended to inspire confidence in the article's veracity.

In addition, a proposed implementation involves the inclusion of a checkbox for authors to assess their past publication history. The proposed feature aims to enable authors to self-report their history of generating dependable and superior content. The study aims to evaluate the reliability and trustworthiness of news articles by taking into account the author's reputation and reliability, providing users with an expanded assessment.

Continuous research, gathering information, and collaborating with trusted sources and respected writers are crucial for the realisation of these future developments. The implementation of these features is a complex task that presents several challenges, such as the requirement for data integration, source confirmation, and author evaluation methods. The aim of this study is to develop a website that provides a reliable and comprehensive tool to combat fake news in the Bangla language. To achieve this goal, we will utilise improvements in natural language processing, web scraping, and data analysis, despite the obstacles that may arise.

Bibliography

- [1] B. Markines, C. Cattuto, and F. Menczer, “Social spam detection,” in *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, Apr. 2009.
- [2] R. Mihalcea and C. Strapparava, “The lie detector: Explorations in the automatic recognition of deceptive language,” in *Proceedings of the ACL-IJCNLP 2009 conference short papers*, 2009, pp. 309–312.
- [3] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, “Who is tweeting on twitter: Human, bot, or cyborg?” In *Proceedings of the 26th annual computer security applications conference*, 2010, pp. 21–30.
- [4] M. R. Mahmud, M. Afrin, M. A. Razzaque, E. Miller, and J. Iwashige, “A rule based bengali stemmer,” in *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2014, pp. 2750–2756.
- [5] N. K. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: Methods for finding fake news,” *Proceedings of the association for information science and technology*, vol. 52, no. 1, pp. 1–4, 2015.
- [6] F. Alam, S. A. Chowdhury, and S. R. H. Noori, “Bidirectional lstms—crfs networks for bangla pos tagging,” in *19th International Conference on Computer and Information Technology (ICCIT), 2016*, IEEE, 2016, pp. 377–382.
- [7] 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.
- [8] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
- [9] W. Wei and X. Wan, “Learning to identify ambiguous and misleading news headlines,” *arXiv preprint arXiv:1705.06031*, 2017.
- [10] M. Aldwairi and A. Alwahedi, “Detecting fake news in social media networks,” *Procedia Computer Science*, vol. 141, pp. 215–222, 2018.
- [11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [12] M. Martinez, “Burned to death because of a rumour on whatsapp,” *BBC News*, 2018.

- [13] A. Ranjan, “Fake news detection using machine learning,” Ph.D. dissertation, 2018.
- [14] C. K. Hiramath and G. Deshpande, “Fake news detection using deep learning techniques,” in *2019 1st International Conference on Advances in Information Technology (ICAIT)*, IEEE, 2019, pp. 411–415.
- [15] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, “Fake news detection using machine learning ensemble methods,” *Complexity*, vol. 2020, pp. 1–11, 2020.
- [16] M. Z. Hossain, M. A. Rahman, M. S. Islam, and S. Kar, “BanFakeNews: A dataset for detecting fake news in Bangla,” English, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 2862–2871, ISBN: 979-10-95546-34-4. [Online]. Available: <https://aclanthology.org/2020.lrec-1.349>.
- [17] M. Zobaer Hossain, M. Ashrafur Rahman, M. Saiful Islam, and S. Kar, “Ban-fakenews: A dataset for detecting fake news in bangla,” *arXiv e-prints*, arXiv–2004, 2020.
- [18] N. F. Baarir and A. Djeflal, “Fake news detection using machine learning,” in *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, IEEE, 2021, pp. 125–130.
- [19] Z. Khanam, B. N. Alwasel, H. Sirafi, and M. Rashid, “Fake news detection using machine learning approaches,” *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012 040, Mar. 2021. DOI: 10.1088/1757-899X/1099/1/012040. [Online]. Available: <https://dx.doi.org/10.1088/1757-899X/1099/1/012040>.
- [20] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, “A comprehensive review on fake news detection with deep learning,” *IEEE Access*, vol. 9, pp. 156 151–156 170, 2021.
- [21] S. Sarker, “Bnlp: Natural language processing toolkit for bengali language,” *arXiv preprint arXiv:2102.00405*, 2021.
- [22] M. Y. Tohabar, N. Nasrah, and A. M. Samir, “Bengali fake news detection using machine learning and effectiveness of sentiment as a feature,” in *2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, IEEE, 2021, pp. 1–8.
- [23] E. Hossain, M. Nadim Kaysar, A. Z. M. Jalal Uddin Joy, M. Mizanur Rahman, and W. Rahman, “A study towards bangla fake news detection using machine learning and deep learning,” in *Sentimental Analysis and Deep Learning: Proceedings of ICSADL 2021*, Springer, 2022, pp. 79–95.
- [24] M. S. A. Chowdhury, A. Hossain, and M. J. Rime, “News literacy in bangladesh,” *Journal of Media Literacy Education*, vol. 15, no. 2, pp. 1–12, 2023.
- [25] A. or Maintainers, *bangla-stemmer*, <https://pypi.org/project/bangla-stemmer/>, Year the package was released.

[26] A. or Maintainers, *bltk*, <https://pypi.org/project/bltk/>, Year the package was released.