

MULTILINGUAL FAKE NEWS DETECTION

BY

Shimul Sutradhar

ID: 191-15-12614

Md. Rahat Islam

ID: 191-15-12895

AND

Tanjima Akhanda Mim

ID:191-15-2455

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

Supervised By

Dr. Sheak Rashed Haider Noori

Professor & Associate Head

Department of CSE

Daffodil International University

Co-Supervised By

Mr. Saiful Islam

Assistant Professor

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY 2023

APPROVAL

This Project/internship titled “**Multilingual Fake News Detection**”, submitted by Shimul Sutradhar, ID No: 191-15-12614, Md. Rahat Islam, ID No: 191-15-12895 and Tanjima Akhanda Mim, ID No: 191-15-2455, to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 02/02/2023.

BOARD OF EXAMINERS



Dr. Touhid Bhuiyan
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



Dr. Sheak Rashed Haider Noori
Professor and Associate Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Sazzadur Ahamed
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Md. Sazzadur Rahman
Associate Professor
Institute of Information Technology
Jahangirnagar University

External Examiner

DECLARATION

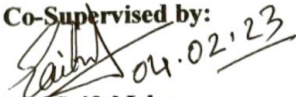
We hereby declare that this project has been done by us under the supervision of **Dr. Sheak Rashed Haider Noori, Professor & Associate Head, Department of CSE** Daffodil International University. We also affirm that neither this project nor any portion of this project has been submitted elsewhere for the purpose of receiving a degree or certificate.

Supervised by:



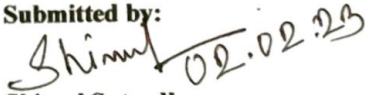
Dr. Sheak Rashed Haider Noori
Professor & Associate Head
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Saiful Islam
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Shimul Sutradhar
ID: 191-15-12614
Department of CSE
Daffodil International University



Md. Rahat Islam
ID: 191-15-12895
Department of CSE
Daffodil International University

A handwritten signature in black ink, written over a diagonal line. The date '01-02-23' is written below the signature.

Tanjima Akhanda Mim

ID: 191-15-2455

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First, we express our deepest appreciation and gratitude to almighty God for His divine blessing, which makes it possible to complete the final year Project/Internship successfully.

We are extremely appreciative and wish to express our profound gratitude to **Dr. Sheak Rashed Haider Noori, Professor & Associate Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine learning and Deep Learning*” to carry out this project. The completion of this project was made possible by his inexhaustible patience, scholarly guidance, continuous encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading numerous subpar drafts and correcting them at every stage.

We would like to express our heartfelt gratitude to **Professor Dr. Touhid Bhuiyan, Professor and Head**, Department of CSE, for his kind help in finishing our project, as well as to other faculty members and the staff of the CSE department at Daffodil International University.

We would like to thank every Daffodil International University classmate who participated in this discussion while completing course work.

Finally, we must respectfully acknowledge the unwavering support and tolerance of our parents.

ABSTRACT

Fake news is becoming more of a concern as the number of people who use the internet and have access to the internet continues to expand on a daily basis. The dissemination of misleading information has the potential to cause harm to individuals. In this research work, we offer a system that can automatically rate the possibility that a news article is fraudulent by using machine learning and deep learning techniques. The system is trained on a multilingual dataset consisting of both fake and real news articles from legitimate news websites and datasets from previous works. To make its predictions, the system uses a combination of natural language processing techniques, machine learning, and deep learning algorithms. We examine the performance of the system using a held-out test set and demonstrate its effectiveness in identifying fake news with a high degree of accuracy. The suggested approach has the potential to be an effective instrument in the battle against fake news, contributing to the reduction of the transmission of false information and preventing readers from being misled. In our work SVM, DT, RF, KNN, Logical Regression, NB, XGBoost, and LSTM are applied models. These models are 93%, 88%, 95%, 88%, 94%, 90%, 91%, and 96% accurate. Among all of the algorithms, LSTM has the highest accuracy of 96%.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of Examiners	i
Declaration	ii
Acknowledgements	iv
Abstract	v
CHAPTERS	
CHAPTER 1 : INTRODUCTION	1-3
1.1 Introduction	01
1.2 Motivation	01
1.3 Rationale of the Study	02
1.4 Research Question	02
1.5 Expected Output	03
1.6 Report Layout	03
CHAPTER 2 : BACKGROUND	4-11
2.1 Terminologies	04
2.2 Related Works	04
2.3 Comparative Analysis and Summary	08
2.4 Scope of the Problem	10
2.5 Challenges	11
CHAPTER 3 : RESEARCH METHODS	12-22
EXPERIMENTAL RESULT AND DISCUSSION LOGY	
3.1 Research Subject and Instrumentation	12
3.2 Data Collection Procedure	12
3.3 Statistical Analysis	14
3.4 Proposed Methodology	14
	vi

3.5 Implementation Requirements	22
CHAPTER 4 : EXPERIMENTAL RESULT AND DISCUSSION	23-26
4.1 Experimental Setup	23
4.2 Experimental Results & Analysis	24
4.3 Discussion	26
CHAPTER 5 : IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	27-29
5.1 Impact on Society	27
5.2 Impact on Environment	28
5.3 Ethical Aspects	28
5.4 Sustainability Plan	29
CHAPTER 6 : SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION AND IMPLICATION FOR FUTURE RESEARCH	30-31
6.1 Summary of the Study	30
6.2 Conclusions	30
6.3 Recommendation	31
6.4 Implication for Further Study	31
REFERENCES	32-33

LIST OF FIGURES

FIGURES	PAGE
Figure 3.1: Proposed Methodology	15
Figure 3.2: Data Preprocessing Steps	17
Figure 4.1: Confusion Matrix of all models	25

LIST OF TABLES

TABLES	PAGE
Table 2.1: Comparison Between Bangla and English Previous fake news detection work.	9
Table 3.1: Sample of Dataset of Fake News.	13
Table 3.2: Statistical Analysis of Dataset.	14
Table 3.3: Raw text data and preprocessed text data.	16
Table 4.1: Model Accuracy.	23
Table 4.2: Classification Report.	26

CHAPTER 1

INTRODUCTION

1.1 Introduction

People are using the internet for communication. Social media has become an integral part of our lives. But the risk of fake news has become a major concern for many users. The proliferation of fake news on social media has become a critical issue for governments, organizations, and individuals around the world. Fake news can cause confusion and spread misinformation, impacting people's lives in a negative way. With the increasing popularity of social media, various techniques are being used to detect and combat the spread of fake news. As such, it is important to develop effective techniques for detecting and filtering out fake news on social media platforms. It doesn't take long for news to spread. There are lots of detection systems that exist. But one of the major problems is that those systems are quite slow. If we look at the last few years, fake news was detected after damage had been done. To decrease the rate of damage, we need a faster system that can detect fake news more effectively and efficiently in a multilingual environment.

One promising approach is the use of classification algorithms for fake news detection on social media. This technique uses a machine learning-based approach to analyze the content of online posts and identify any false information. By leveraging the power of natural language processing and other advanced techniques, classification algorithms can be used to quickly and accurately detect fake news on social media platforms.

1.2 Motivation

In the modern age, every action of a person, a group of people, or everyone relies on information. A single piece of information can make a huge impact on progress or escalate things very quickly. So, it is very important for information like news to be as authentic as possible. Over the last decade, we have come a long way in terms of communication. Passing information through the internet has become a common phenomenon. Publishing news articles on the internet has become normal as the internet has become much more accessible than before.

So, anyone can access information that is available online. This brought about a huge advancement in technology. But at the same time, some risk factors have come to light. Some bad parties use this boon of technology to put themselves in advantageous positions, harming others by publishing fake news that may have some serious, long-lasting, irreversible consequences. In the last few years, we have experienced what the spread of fake news can result in and how much bloodshed has gone on based on fake news. Fake news has the potential of destroying lives and nations.

In the year 2019, eight people have been killed in mob attacks in Bangladesh after false rumors about child abductions spread online. A mother was killed in a mob attack. People suspect her of being a child abductor based on fake news. Those rumors had been detected, but damage had already been done.[29]

Although much progress has been made in the identification of fake news, there is always room for improvement. We aim to enhance false news identification so that such occurrences do not occur in the future.

1.3 Rationale of Study

Nowadays, people interact over the internet. False news may be harmful to an individual's emotional and physical wellbeing. We currently live in a technologically advanced era. We provide a wide range of false news detection tools in both Bengali and English. When someone writes a news report in Bangla or English on the internet, it might be true or false. Sometimes we confuse fake news with real news. We have a tough time determining if a piece of news is real or false. As a result, we developed a dataset and proposed a methodology for detecting fake news. By doing this, we will be able to make people feel more secure when reading internet news. As a consequence of our research, we will discover the finest algorithm for identifying false news in a reliable and effective manner.

1.4 Research Question

Some questions concerning this work arise during the study process. The following are the primary concerns of our work:

- Detecting fake news efficiently.
- Detect fake news accurately.

1.5 Expected Outcomes

- Our main outcome is an automatic fake news detection system.
- To build a robust fake news detection system.
- Make a balanced fake news multilingual dataset.
- To help online platforms(social media) to detect and delete fake news as soon as they are posted online.
- To save people from being attacked online for their profession, religion and race.

1.6 Report Layout

In this report, there are 6 chapters.

- In Chapter 1, we review the overall structure of our study work and divide it into subchapters. For example, the introduction, motivation, reasoning, study subject, and predicted conclusion of our project.
- In Chapter 2, we discussed previous work on false news identification, the scope of the problem, and the challenges in this study.
- In Chapter 3 we will talk about our work procedure, methods and techniques to build a Multilingual Fake News Detection model.
- In Chapter 4, we will go through the Experimental Results and our Build Model Discussion.
- In Chapter 5, we'll talk about how our labor affects society, the environment, ethics, and the long term.
- In Chapter 6, we discussed the work's Summary, Conclusion, and Further Study.

CHAPTER 2

BACKGROUND

2.1 Terminologies

Machine learning is a well-known way of dealing with classification issues. Researchers frequently use machine learning. Machine learning algorithms can learn from previous data and predict future data, which may be used by the researcher to assess behavioral aspects. Image recognition, audio recognition, and facial recognition are just a few of the uses for machine learning. Machine-learning-based technology is becoming increasingly popular. People are currently using social media to express their views and emotions. People gather information from social media. They see news, videos, etc. For our work, we first collect data from an online newspaper called "Kaler Kantho." We discussed terminologies like LSTM. We also discuss terms such as confusion matrix, classification report and hyperparameter tuning.

2.2 Related Works

In the paper author Veronica Perez-Rosas et al. they figure out a fake news detection system. They use two datasets. The first one is the crowdsourced dataset. They take it from different news portals like ABC News, CNN, USA Today, etc. They mixed some fake news from the legitimate news dataset into this one. Another one is taken from a different website. They use SVM classification for computing results. And humans detect fake news to compare results. Their best performance achieves accuracy that is comparable to humans' ability to spot fake news.[2]

One of the most critical problems in NLP is fake news detection. In this paper by Ray Oshikawa, Jing Qian, and William Yang Wang, they survey automated fake news detection from the perspective of NLP. They discuss the pros and cons, potential pitfalls, and drawbacks of each model. They do that based on the following facts: fact-checking, rumor detection, stance detection, and sentiment analysis. They compare different models [2]. Those are Support Vector Machine (SVM), Naive Bayes Classifier (NBC), Logistic Regression (LR), Random Forest Classifier (RFC),

RNN, LSTM, LIAR, FEVER, LSTM, CNN, RST, VSM, GCN, and HC-CB-3. The best performances are given by NLP Shallow Deep (CNN), GCN, and HC-CB-3.[2]

Covering a wide range of topics, the paper by Kai Shuy et al. provides a comprehensive overview of the state of the art in detecting fake news on social media, including: characterizations of fake news based on psychological and social theories; existing algorithms from a data mining perspective; evaluation metrics; and representative datasets. A literature review is conducted, and the information gathered is used to both define the issue of fake news and to provide potential solutions to it. The framework for recognizing fake news across all mediums was built throughout the process of characterizing it. Specifically, they investigated data mining strategies for identifying fake news, such as feature extraction and model construction.[3]

A multi-source multi-class fake news detection framework, MMFD, was proposed by Hamid Karimi, Proteek Chandan Roy, Sari Saba-Sadiya, and Jiliang Tang in their paper. This framework integrates automated feature extraction, multi-source fusion, and automated degrees of fakeness detection into a unified and interpretable model. The efficiency of the suggested framework is shown by experimental findings on real-world data, and more experiments are done to gain insight into the framework's inner workings.[4]

Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad tackled the topic of categorizing false news articles using machine learning models and ensemble approaches in their study. The paper was titled "The Problem of Classifying Fake News Articles." They made use of LR, SVM, KNN, CNN, and LSVM in their analysis. They employed a LIWC tool to extract various textual characteristics from the articles, and they sent the feature set into the models as an input. In order to achieve maximum precision, the learning models underwent training and had their parameters fine-tuned. When compared to other models, several ones have attained much greater levels of accuracy.[5]

In the paper, Kai Shu, Suhang Wang, and Huan Liu discussed how not only the content but also the context mattered in detecting fake news effectively. They proposed building an embedding framework, TriFN, from scratch. The framework will work on the relationship between the publisher, the news, and the user. They

collected the dataset from "FakeNewsNet" and used the data from two social media news platforms. They used information about the tri-relationship as the parameter. It can be seen in the paper that their proposed system brings a higher result than traditional systems, which rely more on content similarity than context.[6]

Z Khanam, B. N. Alwasel, H. Sirafi1, and M. Rashid proposed using the Python scikit-learn module to perform text data tokenization and feature extraction because it includes useful tools such as the Count Vectorizer and Tiff Vectorizer. This approach will result in feature extraction and vectorization. Using Python Sklearn, they divided the dataset into 70% for training and 30% for testing [7]. They applied 7 algorithms: Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbors, Decision Tree, and SVM. They obtained their data from Kaggle, with a 74.5 percent average accuracy. Twitter spam senders were identified using Naive Bayes algorithms, with accuracy ranging from 70% to 71.2%. With a 76% accuracy rate, they tested a variety of algorithms. Their research employs three widely used techniques: Naive Bayes, neural networks, and support vector machines (SVM). The accuracy of Naive Bayes is 96.08% [8]. The accuracy of the neural network and the machine vector (SVM) was 99.9 percent. Using a mixed false message detection model, they used KNN and random forests to produce results that were up to 80% better than the initial ones.

In the paper of Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro, the researchers demonstrated that Facebook posts might be accurately categorized as hoaxes or non-hoaxes based on the users who "liked" them. They applied two classification algorithms, one based on boolean crowdsourcing algorithms and the other on logistic regression. Here, researchers worked with a dataset consisting of 15,500 Facebook posts and 909,236 users and achieved classification accuracy levels of 99% even though the training set only contains a small fraction of the posts (less than 1%). They got 99% for logistic regression and 99.4% for the harmonic algorithm. The performance of logistic regression starts at 91.6% for a training set consisting of 10% of posts and degrades towards 56% for a training set consisting of 0.1% of posts, maintaining a performance margin of 3-4% over harmonic BLC, even though the differences between the logistic regression and harmonic BLC algorithms are not great[8]. They demonstrated the reliability of algorithms by demonstrating their effectiveness while focusing on people

who liked both fake and real postings. These findings imply that charting the information's diffusion pattern can be a useful tool for computerized fake news detection systems.

In the paper by Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar, they publicly release an annotated dataset of 50k Bangla news. They gathered information from popular websites that publish satire news in Bangla, as well as misleading or false context information from www.jaachai.com and www.bdfactcheck.com. They use linear support vector machines (SVM), random forest (RF), and logistic regression. RF scores 55%, SVM scores 91%, and LR scores 53%. The best accuracy is gained by SVM, with 91% accuracy compared to human performance, which is 64.8% of the F1-score[9].

In the paper by Tasnuba Sraboni, Md. Rifat Uddin, Fahim Shahriar, Ruhit Ahmed Rizon, and Shakib Ibna Shameem Pollock, they use a dataset consisting of 51.8k data points. However, 49.6K of them are genuine data, while the remaining 2.3K are fake news. They use passive aggressive classifiers, multinomial naive Bayes, support vector machines, logistic regression, decision trees, and random forests. They have applied the classification algorithms in four ways with the same dataset but with different split train-test models. With the 50:50 train-test model, they got the accuracy of 92.2%, 84%, 92.5%, 91%, 85%, and 91% using passive aggressive classifiers, multinomial naive bayes, support vector machines, logistic regression, decision trees, and random forests, respectively. Their best accuracy is gained by SVM, which has 92.5% accuracy. With the 60:40 train-test model, they got the accuracy of 91.9%, 85%, 91.9%, 90%, 86%, and 91% using passive aggressive classifiers, multinomial naive bayes, support vector machines, logistic regression, decision trees, and random forests, respectively. Their best accuracy is gained by PAC and SVM, with 91.9 percent accuracy. With the 80:20 rain-test model, they got the accuracy of 93%, 86%, 92%, 90%, 86%, and 92% using passive aggressive classifiers, multinomial naive bayes, support vector machines, logistic regression, decision trees, and random forests, respectively. Their best accuracy is gained by PAC, with 93% accuracy. With the 70:30 train-test model, they got the accuracy of 93.7%, 86%, 93.5%, 92%, 86%, and 93.21% using passive aggressive classifiers, multinomial naive bayes, support vector machines, logistic regression, decision trees, and random forests, respectively.

Their best accuracy is gained by PAC, SVM, and RF, with 93% accuracy. With a decreasing test ratio, the accuracy of algorithms is increasing. It is clearly seen that the passive-aggressive classifier (PAC) and support vector machine (SVM) performed well in all scenarios. Other classifiers struggled with different ratios, but 70:30 yielded the best results. [10]

Md. Muzakker Hossain, Zahin Awosaf, Md. Salman Hossan Prottoy, Abu Saleh Muhammod Alvy, and Md. Kishor Morol use a dataset of 50K bangla news in their paper. They have collected the news from various online news portals. They use Logistic Regression (LR), Support Vector Machine (SVM), Multinomial Naive Bayes (MNB), Branch and Bound Algorithm (BNB), Random Forest Classifier (RFC), and Decision Tree Classifier (DTC). The accuracy of the LR, SVM, MNB, BNB, RFC, and DTC is 98%, 98%, 98%, 97%, 99%, and 98%, respectively. Random Forest had the best accuracy of 99% with an F1-score of 0.791.[12]

In the paper by Anika Anjum, Mumenuunessa Keya, Abu Kaisar Mohammad Masum, and Dr. Sheak Rashed Haider Noori, They prepare their dataset with 300 articles. They use 5 types of classification algorithms, such as NB, RF, DT, SVM, and KNN. NB scores 74%, RF scores 82%, DT scores 80%, SVC scores 72%, and KNN scores 70%. With the Random Forest Classifier, they achieved the highest accuracy of 82%.[13]

2.3 Comparative Analysis and Summary

We assessed some previous work related to our task. Because we work with natural language processing, often known as NLP, we linked various papers that deal with both Bangla and English test data. In short, we'll investigate whether machine learning algorithms work extraordinarily well with both Bangla and English text data. In this section, we shall compare one previous attempt to another.

Table 2.1: Comparison Between Bangla and English Previous fake news detection work.

Author	Approach	Best Algorithms Name	Best Accuracy Score
Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, Rada Mihalcea	KNN, SVM, Logistic Regression	SVM	88%
Ray Oshikawa, Jing Qian, William Yang Wang	Support Vector Machine (SVM), Naive Bayes Classifier (NBC), Logistic Regression (LR), Random Forest Classifier (RFC), RNN, LSTM, LIAR, FEVER, LSTM, CNN, RST, VSM, GCN, HC-CB-3	NLP Shallow Deep (CNN), GCN, HC-CB-3	94%
Kai Shuy, Amy Slivaz, Suhang Wangy, Jiliang Tang, and Huan Liuy	LSTM, CNN, RST, VSM	CNN	88%
Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais	LR, SVM, KNN, CNN, and LSVM	SVM	93%
Z Khanam , B N Alwasel , H Sirafil and M Rashid	Random Forest, XGBoost, Naive Bayes, K-nearest Neighbors, Decision Tree, and SVM	SVM	99.9%
Eugenio Tacchini , Gabriele Ballarin , Marco L. Della Vedova , Stefano Moret , and Luca de Alfaro	crowdsourcing algorithms and logistic regression	Logistic Regression	99%
Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar	Linear Support Vector Machine (SVM), Random Forest (RF) and Logistic Regression	Linear Support Vector Machine (SVM)	91%

Tasnuba Sraboni, Md. Rifat Uddin, Fahim Shahriar, Ruhit Ahmed Rizon, Shakib Ibna Shameem Polock	Passive Aggressive Classifier, Multinomial Naive Bayes, Support Vector Machine, Logistic Regression, Decision Tree Classifier, Random Forest	passive-aggressive classifier (PAC) and support vector machine (SVM)	93%
Md. Muzakker Hossain, Zahin Awosaf , Md. Salman Hossan Prottoy, Abu Saleh Muhammad Alvy, and Md. Kishor Morol	Logistic Regression(LR), Support Vector Machine(SVM), Multinomial Naive Bayes(MNB), Branch and bound Algorithm(BNB), Random Forest Classifier(RFC), Decision Tree Classifier(DTC)	Random Forest	99%
Anika Anjum, Mumenunnessa Keya, Abu Kaisar Mohammad Masum, Dr. Sheak Rashed Haider Noori	NB, RF, DT, SVM, KNN. NB	Random Forest	82%

We reviewed previous research on the text classification problem. However, the accompanying table only includes work related to Bangla and English text classification. because it will be really valuable in comparing our proposed model. The table above shows that the majority of the effort is focused on Bangla and English NLP and text classification. SVM had the highest accuracy in our previous study. So, based on previous works, we have a good idea of which algorithms we will use for our project.

2.4 Scope of the Problem

False news detection requires a significant amount of work in a given language. However, only a modest amount of work in multiple languages has been done to detect fake news. Our collected dataset contains only current data. We begin by preprocessing the input with NLP before applying machine learning and deep learning

models. Finally, we determined that this dataset delivers the best results for best accuracy, so we are using it to determine best accuracy.

2.5 Challenges

There are some challenges we have encountered during this work. The primary challenge was gathering the data set. Our main target is to create a balanced data set, which will help us get maximum accuracy. At the beginning, write a script for collecting data from online news sources. Many newspapers are already blocked from scraping. At first, we had to check which Bangladeshi newspapers allow scraping. Kaler Kantho allowed scraping. So we run scraping at Kaler Kantho and generate raw data. These records must now be classified. We read the data and manually classify it for classification. In that dataset, there are lots of unwanted things like extra white space, null values, URLs, etc. At preprocessing, we need to remove those things.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

We want to implement a system that is based on a model and has the ability to detect if a news article is a fake one or not. We have to compile a dataset to feed the model and understand what we are doing with this work. The model will classify a news article based on its fakeness. So, only whether a news article is fake or not will matter. So, there will be two classes: those that are fake or not fake, making the work a binary classification. We can approach it in two different ways while building a machine learning model. One is supervised learning, and the other is unsupervised learning. In supervised learning, we classify some data and feed it into the system so that it gains the ability to determine the classes of the data to be predicted based on the fed dataset analysis. But in the unsupervised learning method, the model is fed unclassified data, and the machine clusters that data, finding patterns from the dataset. To implement our system, we collected our data from different sources, verified them, and fed the model a classified dataset. So, our approach here is a supervised learning approach. As the problem we will be solving is a binary classification problem, we will be working with some classification algorithms. Here is our collected data, and the data to be checked after the system is complete is text data. So, we will be focusing on those classification algorithms that work best with text data in terms of performance and accuracy. Some of these algorithms are the support vector machine (SVM), decision tree (DT), logistic regression (LR), and random forest (RF) algorithms. In the proposed methodology section, all of them will be discussed, including their work procedure and their performance with our provided dataset. We will evaluate the model based on precision and recall. The evaluation methods will also be discussed in the proposed methodology section.

3.2 Data Collection Procedure

When preparing a dataset, if the data is more balanced and reliable, the machine learning algorithms will perform better, thus training the machine perfectly. Data collection is a crucial matter for that reason. So, we have to build our dataset while respecting balance and reliability. We collected data from an online news portal

named "Kaler Kantho." "Kaler Kantho" is a very popular daily news publisher organization in Bangladesh. Most of the news articles are not fake news from this source. But there were some news articles that were verified to be fake and labeled fake manually. This was about the Bangla news portion. But we want our work to be multilingual too. So, for the English language, there has been some work done before. So, because there were verified fake news datasets on Kaggle, FakeNewsNet[27] for English and BanFakeNews[28] for Bangla, creating an English language dataset wasn't too difficult. Kaggle is a website where many research-related resources are made open-source for the betterment of science and technology. We collected the data, created the dataset, and then generalized, merged, and shuffled the data. Our final dataset has 11,120 articles. Table 3.1 provides a portion of the collected dataset.

Table 3.1: Sample of Dataset of Fake News

News	Class
গাজীপুরে ২৯ বছর আগের একটি চাঞ্চল্যকার খুনের মামলার রায়ে নিহতের দুই ভাইকে আমৃত্যু কারাদণ্ড দেওয়া হয়েছে। এ ছাড়া অন্য এক ভাই এবং চার প্রতিবেশী ও ভাড়াটে খুনিকে যাবজ্জীবন কারাদণ্ড দেওয়া হয়েছে। আজ সোমবার গাজীপুরের জেলা ও দায়রা জজ দ্বিতীয় আদালতের বিচারক বাহাউদ্দিন কাজী এই রায় দেন। আলোচিত এই মামলার বাদী ছিলেন নিহত ও সাজাপ্রাপ্তদের ভাই। অন্যদিকে মামলায়	Not Fake
পাবলিক পরীক্ষার প্রশ্ন ফাঁস সংক্রান্ত আলাদা ৯টি মামলার রায় দিয়েছেন রাজশাহীর সাইবার ট্রাইব্যুনাল আদালত। রায়ে আসামিদের সর্বোচ্চ সাত বছরসহ বিভিন্ন মেয়াদে কারাদণ্ডের আদেশ দেওয়া হয়েছে। এ ছাড়া রায়ে এক আসামি খালাস পেয়েছেন। আজ সোমবার রাজশাহীর সাইবার ক্রাইম ট্রাইব্যুনালের বিচারক মো. জি.....	Fake
The United States would prefer that Nobel Peace Prize recipient and human rights activist Liu Xiaobo get cancer treatment outside of China, according to Terry Branstad, the newly appointed U.S. Ambassador to China, who made the statement on Wednesday. Liu, then 61 years old, was tried and convicted of encouraging subversion of state power in 2009, receiving a sentence of 11 years in jail.....	Not fake
With the advent of print on demand and e-readers, parents of toddlers have a fantastic tool to help keep their little ones engaged in reading. They have customizable books, featuring their child as the protagonist. There must be	Fake

some parents of toddlers in the Trump administration. They have discovered that the only way to get their boss to read a security briefing is to treat him just like a toddler.....	
---	--

3.3 Statistical Analysis

After collecting the data, our dataset has a total of 10104 news articles classified as "fake" or "not fake." From all the data, 5683 are "not fake" and 4421 are "fake." The data has news articles in two languages, which are Bengali and English. There are 3566 Bengali and 6538 English news articles. There are 1300 Bengali "fake news," 3121 English "fake news," 2266 Bangla "not fake news," and 3417 English "not fake news." Table 3.2 will provide a detailed statistical analysis.

Table 3.2: Statistical Analysis of Dataset

Languages	Percentage	Fake	Percentage	Not fake	Percentage
Bengali	35.29%	1300	12.86%	2266	22.43%
English	64.71%	3121	30.89%	3417	33.82%
Total	100.00%	4421	43.75%	5683	56.25%

3.4 Proposed Methodology

Our research methodology is going to be discussed in this section. We tried six supervised learning classifiers. They are SVM (Support Vector Machine), DT (Decision Tree), RF (Random Forest), KNN (K Nearest Neighbor), Logistic Regression, NB (Naive Bayes), XGB (Extreme Gradient Boosting), and LSTM. We created our own dataset and used these classification methods on it. Although it can be challenging to get the right resources for Bangla, we make every effort to ensure that our work is accurate. We separated our labor into a few steps to accomplish this. The steps of our methodology are shown in Figure 3.1. In Section 3.2, we discussed the data collection process. The remaining steps of the methodology are given on the next page.

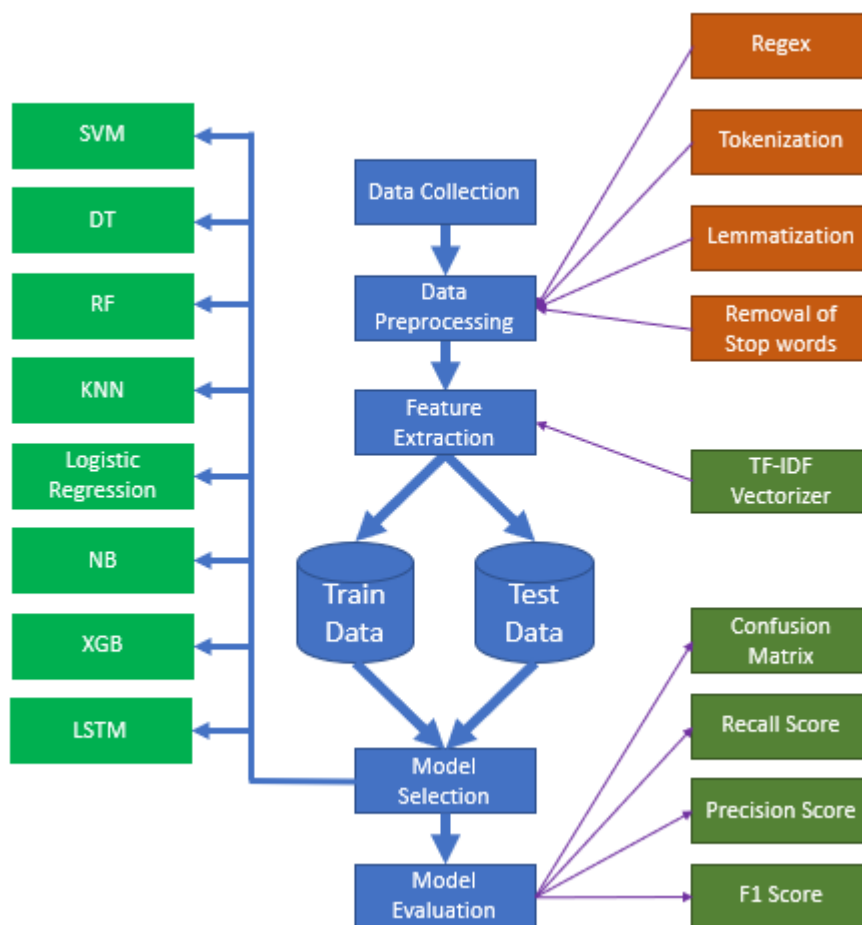


Figure 3.1: Proposed Methodology

3.4.1 Data Preprocessing

In the data collection period, we collected data from various sources. The data scraped from news portal websites had some issues. which was that some elements of the website were embedded in the text data, and some urls were also in the text data from the web scraper. There are also a huge number of special characters in the raw data. This data will behave unusually if fed into the classifier model. They were fixed through the use of scripts. Only English characters were made present in the English text data, and only Bangla characters were made present in the Bangla text. This was done with the help of regex, tweaking it using unicode ranges ("`[u0980-u09FF]`") for Bengali characters, and English characters were handled normally. Some contractions were removed from both languages' text. Table 3.3 shows a portion of the dataset after preprocessing.

Table 3.3: Raw text data and preprocessed text data

Raw Text Data	Preprocessed Text Data
<p>গাজীপুরে ২৯ বছর আগের একটি চাঞ্চল্যকার খুনের মামলার রায়ে নিহতের দুই ভাইকে আমৃত্যু কারাদণ্ড দেওয়া হয়েছে। এ ছাড়া অন্য এক ভাই এবং চার প্রতিবেশী ও ভাড়াটে খুনিকে যাবজ্জীবন কারাদণ্ড দেওয়া হয়েছে। আজ সোমবার গাজীপুরের জেলা ও দায়রা জজ দ্বিতীয় আদালতের বিচারক বাহাউদ্দিন কাজী এই রায় দেন। আলোচিত এই মামলার বাদী ছিলেন নিহত ও সাজাপ্রাপ্তদের ভাই। অন্যদিকে মামলায়</p>	<p>গাজীপুরে ২৯ বছর আগের চাঞ্চল্যকার খুনের মামলার রায়ে নিহতের ভাইকে আমৃত্যু কারাদণ্ড এক ভাই প্রতিবেশী ভাড়াটে খুনিকে যাবজ্জীবন কারাদণ্ড সোমবার গাজীপুরের জেলা দায়রা জজ দ্বিতীয় আদালতের বিচারক বাহাউদ্দিন কাজী রায় আলোচিত মামলার বাদী নিহত সাজাপ্রাপ্তদের ভাই অন্যদিকে মামলায়</p>
<p>পাবলিক পরীক্ষার প্রশ্ন ফাঁস সংক্রান্ত আলাদা ৯টি মামলার রায় দিয়েছেন রাজশাহীর সাইবার ট্রাইব্যুনাল আদালত। রায়ে আসামিদের সর্বোচ্চ সাত বছরসহ বিভিন্ন মেয়াদে কারাদণ্ডের আদেশ দেওয়া হয়েছে। এ ছাড়া রায়ে এক আসামি খালাস পেয়েছেন। আজ সোমবার রাজশাহীর সাইবার ক্রাইম ট্রাইব্যুনালের বিচারক মো. জি.....</p>	<p>পাবলিক পরীক্ষার প্রশ্ন ফাঁস সংক্রান্ত আলাদা ৯টি মামলার রায় দিয়েছেন রাজশাহীর সাইবার ট্রাইব্যুনাল আদালত রায়ে আসামিদের সর্বোচ্চ সাত বছরসহ মেয়াদে কারাদণ্ডের আদেশ রায়ে এক আসামি খালাস পেয়েছেন সোমবার রাজশাহীর সাইবার ক্রাইম ট্রাইব্যুনালের বিচারক মো জি.....</p>
<p>Newly appointed U.S. Ambassador to China Terry Branstad said on Wednesday the United States would like to see Nobel Peace Prize-winning activist Liu Xiaobo treated elsewhere for cancer, and that the two countries must work together on human rights. Liu, 61, was jailed for 11 years in 2009 for inciting subversion of state power after he helped write a.....</p>	<p>newly appoint ambassador china terry say united state would like see peace prize win activist treat elsewhere cancer two country must work together human right jail year incite subversion state power help write</p>
<p>With the advent of print on demand and e-readers, parents of toddlers have a fantastic tool to help keep their little ones engaged in reading. They have customizable books, featuring their child as the protagonist. There must be some parents of toddlers in the Trump administration. They have discovered that the only way to get their boss to read a security briefing is to treat him just like a toddler.....</p>	<p>print demand e reader parent toddler fantastic tool help keep little one engage read book feature child protagonist must parent toddler trump administration discover way get read security briefing treat like toddler</p>
<p>রাজশাহী মেডিক্যাল কলেজ হাসপাতালে রাবি</p>	<p>রাজশাহী মেডিক্যাল কলেজ</p>

<p>শিক্ষার্থীদের ওপর হামলার প্রতিবাদ জানিয়ে প্রতীকি অনশনে বসেছেন রাবি অধ্যাপক মো. ফরিদ উদ্দিন খান। সোমবার (৩১ অক্টোবর) সকাল ১০টা থেকে বিকেল ৫টা পর্যন্ত বিশ্ববিদ্যালয়ের প্রশাসনিক ভবনের সামনে জোহা চত্বরের.....</p>	<p>হাসপাতালে রাবি শিক্ষার্থীদের ওপর হামলার প্রতিবাদ জানিয়ে প্রতীকি অনশনে বসেছেন রাবি অধ্যাপক মো ফরিদ উদ্দিন খান সোমবার ৩১ অক্টোবর সকাল ১০টা বিকেল ৫টা বিশ্ববিদ্যালয়ের প্রশাসনিক ভবনের জোহা চত্বরের</p>
---	--

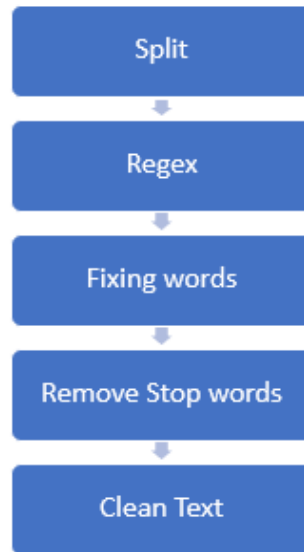


Figure 3.2: Data Preprocessing Steps

3.4.2 Feature Extraction

We have clean text after preprocessing our data, but we can't use this text in our machine learning model. But we need to get the right features, because how well the machine learning model works depends on getting the right features from the text. We need to pull the features out of the clean text. For feature extraction, the text data must be turned into numerical values. We must run feature extraction in a proper way so that it impacts the machine learning model in a positive way in terms of performance. To do this, we need to use some methods that turn our text data into vectors, which are numbers. This is what we call "one-hot encoding." TF-IDF is a common and advanced way to pull out features from processed text data. Using TF-IDF vectorizer methods to pull out features can sometimes make the proposed model more accurate. because it uses the weight of each word in the whole document to make a matrix. The formula behind it is:

$$tf - idf(t) = tf(t, d) * idf(t) \quad \dots \dots \dots (i)$$

$$tf(t, d) = \sum_{x \in d} fr(x, t) \quad \dots \dots \dots (ii)$$

$$idf(t) = \log \frac{|D|}{1 + |d : t \in d|} \quad \dots \dots \dots (iii)$$

Here

t and d are terms.

$fr(x, t)$ is frequency count of term

$|D|$ is the number of whole documents.

$|d : t \in d|$ is the number of document where t appears.

3.4.3 Model Selection

In machine learning techniques, there are two types of learning. Learning Techniques: Supervised and Unsupervised In our work, we have a dataset to train a model. That is why our model's learning method will be supervised. We'll be employing some supervised classifier algorithms. We apply eight different algorithms to our dataset. They are SVM (Support Vector Machine), DT (Decision Tree), RF (Random Forest), KNN (K Nearest Neighbor), Logistic Regression, NB (Naive Bayes), XGB (Extreme Gradient Boosting), and LSTM. We divided our data into two parts after vectorizing it: train data and test data. Here, 70% is training data and 30% is test data. Eight different classifier models will be worked on using our training data, and the testing data will evaluate the model. On our dataset, support vector machines and random forest classifiers perform very well. The algorithms and their performance, based on our Bangla fake news dataset, are described below.

3.4.3.1 Support Vector Machine

The SVM algorithm is based on the layout of every data point in a set of dimensions (the number of accessible features), with the number of specified coordinates determining the value of each attribute. The SVM algorithm plots a data item in n-dimensional space using the coordinates that correspond to the values of each feature in a set of n features. The data is classified using the hyper-plane that was

created to divide the two classes. We got an accuracy of 93% using this algorithm. The confusion matrix of SVM on our dataset is [1175 , 139], [56, 1662].

3.4.3.2 Decision Tree

In the realm of classification, the decision tree is a crucial technique that uses a structure very similar to a flowchart. Attribute conditions are "tested" at each internal node in the decision tree, and the tree branches accordingly. After calculating all attributes, a class label is eventually added to the leaf node. The rule of categorization is shown by the length of a line drawn from the plant's base to its tip. The fact that it may be used with a category and a dependent variable They do a fantastic job of highlighting key factors and providing clear examples of interrelationships. They are crucial in creating novel variables and characteristics that aid in data exploration and provide reliable predictions of the sought-after variable. Tree-based learning algorithms are often used in the context of predictive modeling, where supervised learning approaches are used to construct high accuracy. Using this technique, we were able to achieve an accuracy of 88%. The confusion matrix using this classifier is [1156, 158], [184, 1534].

3.4.3.3 Random Forest

Random Forest is based on the idea of generating many decision tree algorithms, each of which yields a unique result. Random forest employs the results predicted by many decision trees. Random forest selects randomly a subcategory of qualities from each group to guarantee that the decision trees are distinct from one another. Random forest works best when paired with non-correlational decision trees[16]. By using this method, we achieved a 95% accuracy. The confusion matrix using this classifier is [1213, 101], [41, 1677].

3.4.3.4 K-Nearest Neighbor

The K-nearest neighbor technique is one of the most used classification algorithms. Moreover, it is applicable to regression problems. It measures the distance between the proper outcome, or dependent variable, and one or more independent factors (our features). Utilizing the Euclidean distance formula, the distance is calculated. This

method combines data points that are similar or closer to the projected outcome. Based on the value of k (the number of neighbors), it indicates how much data was necessary to form a group. In this instance, $k = 3$ is applied. This algorithm is incapable of predicting the outcome since it memorizes the formed groupings and compares test results to those groups[19]. This method resulted in an accuracy of 88%. Using this classifier, the confusion matrix is [1074, 240], [108, 1610].

3.4.3.5 Logistic Regression

A logistic algorithm estimates the probability of an event occurring. Likely, this event will happen or not based on the given dataset's dependent variable. It also estimates the relationship between a dependent variable and an independent variable. We use binary logistic regression because our data set has a binary class[25]. We got an accuracy of 94% using this algorithm. The confusion matrix using this classifier is [1200, 114], [56, 1662].

3.4.3.6 Naïve Bayes

Also known as a basic probabilistic classifier, Naive Bayes is a classification algorithm. In its simplest form, this classifier is a collection of Bayesian-based classification algorithms. Instead of being a solitary method, this classifier is a collection of well-known classification algorithms that all work on the same idea. We used the multinomial Naive Bayes classifier into our study. Mainly because multinomial naive Bayes works so well with text document data. Documents of some kind are relevant to our work[20]. The accuracy of this method was best in our dataset. That algorithm accurately anticipated how we would be divided into groups. The accuracy is 90% with the confusion matrix [1074, 240], [45, 1673].

3.4.3.7 Extreme Gradient Boosting Classifier (XGB)

To put it simply, XGB is a machine learning method that employs gradient boosting techniques and is based on decision trees. When the data is unstructured, like text data, picture data, etc., this method makes precise predictions. Common applications of this approach include the resolution of classification and regression issues. Our categorization task was successfully predicted by this algorithm[26]. This classifier

predicted pretty much good accuracy 91% with a confusion matrix [1166, 148], [119, 1599].

3.4.3.8 Long Short-Term Memory (LSTM)

When it comes to learning sequence prediction problems with long-term dependencies, long short-term memory (LSTM) RNNs excel. However, LSTM's feedback connections allow it to analyze the whole data sequence, unlike single data points like images. There are a number of potential applications for this, including automated translation and speech recognition. LSTMs, or Long Short-Term Memories, function well in a wide variety of tasks. [24]. This classifier predicted pretty good accuracy, about 96%, with the confusion matrix [1256, 63], [54, 1659].

3.4.4 Model Evaluation

Only by comparing training and testing outcomes can we determine our model's efficacy. Our model's performance may be judged by considering the following information. The results of our model should only be trusted once they have been subjected to cross-validation. Our model's efficacy may then be evaluated by generating a classification report. The next sections will provide a more in-depth explanation.

3.4.4.1 Classification Report

We can't declare this model is the best for this data set without considering the cross-validation score. Moreover, we need to assess a wide range of model parameters. The results and interpretations discussion includes extensive discussion of the parameters.

3.4.4.1.1 Confusion Matrix

It is a table used to show the performance of a machine learning model based on a collection of test output data [12]. It analyzes performance by computing four terms: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). This will be briefly described in the experiment and results section.

3.4.4.1.2 Precision Score

Precision is defined as the percentage of correct predictions made relative to the total number of correct forecasts made. Positive predictive value or PVV is another name for this.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad \dots \dots \dots \text{(iv)}$$

3.4.4.1.3 Recall Score

Calculate t as the fraction of correct predictions that were made out of a total of.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad \dots \dots \dots \text{(iiv)}$$

3.5 Implementation Requirements

For implementing our work Google colab is enough. Python libraries needed are given below.

- Python 3.8
- Pandas
- NLTK
- Regular Expression
- Scikit-Learn
- NumPy

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experimental Setup

The title of our research project is "Multilingual Fake News Detection." We employ machine learning and deep learning techniques in our work. In this section, we will explain our implemented model, which we tested on our supplied dataset of bogus news. The performance of almost every algorithm is quite outstanding. However, some of them provide the finest performances. After applying machine learning and deep learning algorithms to raw data, the accuracy is below 30%. When we analyze data, our precision increases. In the methodology, we outline our preprocessing procedures. After preprocessing, models are applied. SVM, DT, RF, KNN, Logical Regression, NB, XGBoost, and LSTM are applied models. These models are 93%, 87%, 95%, 88%, 94%, 90%, 90%, and 96% accurate. These models perform quite well on our preprocessing data. Table 4.1 displays every model's accuracy depending on each model's degree of precision.

Table 4.1: Model Accuracy

Model Name	Accuracy
SVM	93%
DT	88%
RF	95%
KNN	88%
LR	94%
NB	90%
XGBoost	90%
LSTM	96%

We can't view our model as a great model for our data set. We create a confusion matrix, a categorization report, accuracy, recall, average accuracy, and average recall.

4.2 Experimental Results and Analysis

Machine learning models cannot predict everything with 100 percent accuracy. We must make an effort to find an ideal model solution. In our study, we experiment with our models, utilizing strategies such as hyperparameter tuning. Sometimes, hyperparameter tuning assists us in identifying the right model parameters for our dataset. Therefore, we chose to tweak the parameters of our models depending on our dataset. Then, tune hyperparameters with the assistance of a Python library. After tuning our models based on our dataset, we see that the accuracy of certain methods increased marginally while that of others decreased. After adjusting our models, SVM, DT, RF, KNN, Logical Regression, NB, XGBoost, and LSTM models achieved respective accuracy scores of 93%, 88%, 95%, 88%, 94%, 90%, 91%, and 96%. We see that there is no difference between before and after parameter adjustments in the XGB classifier. Furthermore, the precision of the DT and RF models decreases during hyperparameter tuning. We thus maintain the default settings for the DT, RF, and XGB classifiers. In table 4.2, we provide the parameters of every model utilized in our research. In our study, the hyperparameter tweaking experiment yielded a high degree of precision. However, based just on the accuracy score, we cannot determine which model performs best with our dataset. We test our model using measures such as the accuracy score, the recall score, the average precision, and the average recall score. Initially, we created a confusion matrix using our dataset and each of the categorization models we stated. The depiction in the image depicts the confusion matrix for all models. Then, with the confusion matrix

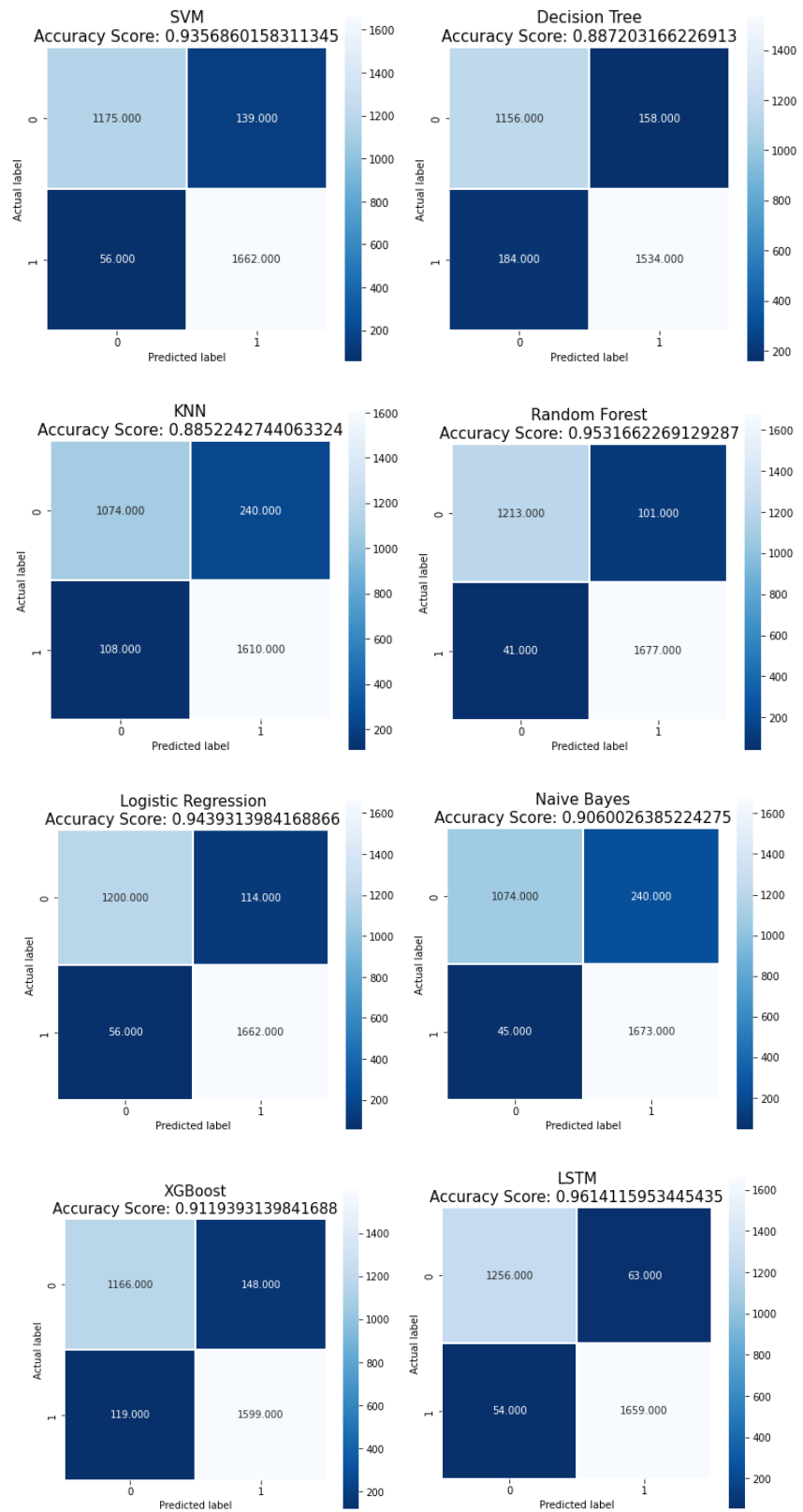


Figure 4.1: Confusion Matrix of all models

Table 4.2: Classification Report

Algorithm Name	Class	Precision	Recall	Average Precision	Average Recall	Accuracy Score
SVM	True	0.95	0.89	0.93	0.93	0.93
	False	0.92	0.96			
Decision Tree	True	0.86	0.87	0.88	0.88	0.88
	False	0.90	0.89			
KNN	True	0.90	0.81	0.88	0.87	0.88
	False	0.87	0.93			
Random Forest	True	0.96	0.92	0.95	0.94	0.95
	False	0.94	0.97			
Logistic Regression	True	0.95	0.91	0.94	0.94	0.94
	False	0.93	0.96			
Naive Bayes	True	0.95	0.81	0.91	0.89	0.90
	False	0.87	0.97			
XGBoost	True	0.90	0.88	0.91	0.90	0.91
	False	0.91	0.93			
LSTM	True	0.95	0.95	0.96	0.96	0.96
	False	0.96	0.96			

4.3 Discussion

In the end, we are trying to contribute to the multilingual (Bangla and English) research domain. Worldwide, there has been lots of work in different languages. We were able to create a machine learning model that can accurately detect multilingual fake news in this work. We use eight different algorithms in our model. Among all of the algorithms, LSTM, logistic regression and random forest detect multilingual fake news accurately with the highest accuracy of 96%.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Every day, the number of individuals who have access to and utilize the internet increases, and as a result, the prevalence of fake news increases. False information has the capacity to cause harm to individuals. In this research work, we propose a system that uses machine learning and deep learning techniques to evaluate automatically the possibility that a news article is fraudulent. The system is trained using a multilingual dataset of both fake and actual news stories from authentic news websites and datasets from earlier efforts, and makes predictions using a combination of natural language processing techniques, machine learning, and deep learning algorithms. We examine the performance of the system on a held-out test set and demonstrate its usefulness in recognizing bogus news with high precision. The suggested method has the potential to be a valuable instrument in the fight against fake news, helping to limit its spread and safeguarding readers from being deceived.

- Fake news can have an effect on the real world, like when it spreads false information about a public health issue or causes panic. During the COVID-19 pandemic, for example, fake news about the virus and how to treat it caused confusion and put public safety at risk.
- Fake news can also change the way people talk and how politics work. It can change what people think and how they feel, and it can even change how elections turnout.
- Spreading fake news can be bad for people who are already at a disadvantage because it can reinforce harmful stereotypes and spread discrimination.
- Fake news can hurt the credibility of real journalism and make it harder for people to tell the difference between fact and fiction. This can make people less trusting of the media and other institutions, which can have serious long-term effects.

Before we share something, we should be careful about where we get our information and check the facts. This can make it harder for bad news and fake news to spread and hurt people.

5.2 Impact on Environment

False news and erroneous information may be harmful to the environment. Spreading misleading information regarding the safety or efficacy of certain environmental policies or technology, for instance, might hinder attempts to tackle critical environmental issues and safeguard natural resources. Similarly, false news may affect the environment if it pushes people to consume dangerous goods or engage in harmful activities.

It is crucial to consider how our activities, such as sharing knowledge, may impact the environment. Additionally, we should use credible sources and fact-check material to ensure that we are not disseminating false or misleading information.

5.3 Ethical Aspects

The dissemination of false news may have grave ethical repercussions, since it can erode faith in media and institutions, promote confusion and strife, and even inspire violence. It may also have detrimental effects on people who are exposed to erroneous or deceptive information. The potential for damage to individuals and society is one of the primary ethical problems associated with fake news. Fake news may disseminate disinformation, promote undesirable ideas or behaviors, and be used as a weapon to influence public opinion and interfere with elections and other democratic processes. The problem of accountability is a further ethical concern. Those who generate and distribute fake news may do so anonymously or under false names, making it difficult to hold them responsible for the damage they do.

The production and dissemination of false news violates the ethical standards of honesty and integrity, and it may have severe implications for people and society as a whole. It is essential to be cautious in spotting and refuting false news and to encourage ethical and responsible communication.

5.4 Sustainability Plan

The work will be sustainable if it helps the next generation avoid fake news as quickly and accurately as possible. This is why we made the model, and it's also why we need to fix it right so that people can find the right answer quickly. We should have known about this idea, and we need to learn more about it and fully comprehend it. So, we need to keep this model in good shape.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

Our work is related to multilingual NLP. In research work, datasets are important for executing the work. In this work, we are making a machine learning model for detecting fake news in a multilingual environment. This model is very useful for detecting fake news. So, while working on this model, we encounter a number of issues and challenges. At first, we plan for this project; after that, we formulate problems, collect data, preprocess data, train and test data, select a model, and conduct an evaluation. After executing all of those steps, we were finally able to build our model. Working with our native language is a source of pride for us, and in the world-wide domain, we recognize our language easily.

6.1 Conclusion

The research conducted on multilingual fake news detection has shown that it is a complex and multifaceted problem that requires the use of multiple approaches and techniques to be effectively addressed. By analyzing a large dataset of news articles in various languages, we were able to identify several key characteristics that are common among fake news articles and develop machine learning models that can accurately detect fake news with high precision and recall.

Overall, our findings suggest that the use of natural language processing techniques, combined with the incorporation of contextual information and knowledge about the credibility of sources, can significantly improve the accuracy of fake news detection models. While further research is needed to fully understand and address the issue of fake news in a multilingual context, our work represents an important step towards building reliable and robust systems for detecting and combating the spread of fake news online.

6.3 Recommendation

We have many recommendations for our work. In this section, we will increase our dataset in order to improve the precision of our model. We apply classification techniques based on supervised machine learning in our work. And just one vectorization technique is used in the text data transformation section. There are several techniques and methods for evaluating massive datasets. As a result, the model and techniques will predict the detection of multilingual fake news with more accuracy.

6.4 Implication for Further Study

Our work has certain limits and disadvantages. In our model, for instance, we exclusively use machine learning methods. Other than this, only TDF Vectorized text processing algorithms are used. Also, our data is not adequate. Thus, we increase our data count. We want to use deep learning techniques such as SVM, LSTM, etc. on our dataset. Without expanding the data amount, the accuracy of the deep learning model cannot be improved. In addition, we use Word Embedding algorithms to vectorize the text input into numeric values. This study achieves 95% accuracy, but we cannot stop here; in the future, we will make every effort to increase this dataset's precision.

Reference

- [1] Pérez-Rosas V, Kleinberg B, Lefevre A, Mihalcea R. Automatic detection of fake news. arXiv preprint arXiv:1708.07104. 2017 Aug 23.
- [2] Oshikawa R, Qian J, Wang WY. A survey on natural language processing for fake news detection. arXiv preprint arXiv:1811.00770. 2018 Nov 2.
- [3] Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H., 2017. Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter, 19(1), pp.22-36.
- [4] Karimi, H., Roy, P., Saba-Sadiya, S. and Tang, J., 2018, August. Multi-source multi-class fake news detection. In Proceedings of the 27th international conference on computational linguistics (pp. 1546-1557), Aug 2018.
- [5] Ahmad, I., Yousaf, M., Yousaf, S. and Ahmad, M.O., 2020. Fake news detection using machine learning ensemble methods. Complexity, 2020.
- [6] Shu, Kai, Suhang Wang, and Huan Liu. "Beyond news contents: The role of social context for fake news detection." Proceedings of the twelfth ACM international conference on web search and data mining. Jan 2019.
- [7] Mukhaini, F.A., Abdoulie, S.A., Kharuosi, A.A., Ahmad, A.E., & Aldwairi, M. (2022). FALSE: Fake News Automatic and Lightweight Solution. 2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), 49-54, Jul 2022.
- [8] Khanam, Z., et al. "Fake news detection using machine learning approaches." IOP Conference Series: Materials Science and Engineering. Vol. 1099. No. 1. IOP Publishing, Mar 2021.
- [9] Tacchini, Eugenio, et al. "Some like it hoax: Automated fake news detection in social networks." Proceedings of the Second Workshop on Data Science for Social Good, Volume 1960, Apr 2017.
- [10] Hossain, Md Zobaer, et al. "Banfakenews: A dataset for detecting fake news in bangla." arXiv preprint arXiv:2004.08789 (2020), Apr 2022.
- [11] Sraboni T, Uddin M, Shahriar F, Rizon RA, Pollock SI. FakeDetect: Bangla fake news detection model based on different machine learning classifiers (Doctoral dissertation, Brac University), 2021.
- [12] Hossain, Md, et al. "Approaches for Improving the Performance of Fake News Detection in Bangla: Imbalance Handling and Model Stacking." Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021. Springer, Singapore, 2022.
- [13] Anjum, Anika, et al. "Fake and Authentic News Detection Using Social Data Strivings." 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, Jul 2021.
- [14] Faigman, David L., and A. J. Baglioni. "Bayes' theorem in the trial process." Law and Human Behavior 12.1 (1988): 1-17.
- [15] Aldwairi, Monther, and Ali Alwahedi. "Detecting fake news in social media networks." Procedia Computer Science 141 (2018): 215-222, 2018.
- [16] R. Darnton, The True History of Fake News, The New York Review of Books, 2017.
- [17] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election", The Journal of Economic Perspectives, vol. 31, no. 2, pp. 211-235, 2017.

- [18] J. Gorbach, "Not Your Grandpa's Hoax: A Comparative History of Fake News", *American Journalism*, vol. 35, no. 2, pp. 236-249, 2018.
- [19] M. Luo, J.T. Hancock and D.M. Markowitz, "Credibility Perceptions and Detection Accuracy of Fake News Headlines on Social Media: Effects of Truth-Bias and Endorsement Cues", *Communication Research*, vol. 49, no. 2, pp. 171-195, 2022.
- [20] S.K. Sharma, S. Kumar, M.S. Manral and A. Verma, "Paid News Syndrome in Print Media: A Study Based on Selective Newspapers Readers in Jaipur City", *Journal of Positive School Psychology*, vol. 6, no. 2, 2022.
- [21] N. Alsharif, "Fake Opinion Detection in an E-Commerce Business Based on a Long-Short Memory Algorithm", *Soft Comput.*, 2022.
- [22] A.R. DiMaggio, "Conspiracy Theories and the Manufacture of Dissent: QAnon the 'Big Lie' Covid-19 and the Rise of Rightwing Propaganda", *Critical Sociology*, 2022.
- [23] A. Khalil, M. Jarrah, M. Aldwairi and Y. Jararweh, "Detecting Arabic Fake News Using Machine Learning", In *Proceedings of the Second International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pp. 171-177, 15-17 Nov. 2021.
- [24] N. Conroy, V. Rubin and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News", In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, vol. 1, no. 1, pp. 85, 2015.
- [25] KA. Stevens, S. Chengjie, L. Bingquan and W. Xiaolong, "Transfer Learning and GRU-CRF Augmentation for COVID-19 Fake News Detection", *Computer Science and Information Systems Mathematics*, vol. 10, no. 4, pp. 585, 2022.
- [26] S. Ahmed, K. Hinkelmann and F. Corradini, "Fact Checking: An Automatic End to End Fact Checking System", *Combating Fake News with Computational Intelligence Techniques. Studies in Computational Intelligence*, vol. 1001, 2022.
- [27] Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media." *Big data* 8, no. 3 (2020): 171-188.
- [28] Hossain MZ, Rahman MA, Islam MS, Kar S. Banfakenews: A dataset for detecting fake news in bangla. arXiv preprint arXiv:2004.08789. 2020 Apr 19.

Websites:

- [29] BBC, available at << <https://www.bbc.com/news/world-asia-49102074>>>, last accessed on 21/5/2022 at 9.56 PM.
- [30] Kaler Kantho, available at << <https://www.kalerkantho.com/>>>, last accessed on 29/12/2022 at 10.59 PM.

MULTILINGUAL FAKE NEWS DETECTION

ORIGINALITY REPORT

24%

SIMILARITY INDEX

21%

INTERNET SOURCES

10%

PUBLICATIONS

10%

STUDENT PAPERS

PRIMARY SOURCES

1 dspace.daffodilvarsity.edu.bd:8080 11%
Internet Source

2 Submitted to Daffodil International University 4%
Student Paper

3 www.financialexpress.com 1%
Internet Source

4 Submitted to Bournemouth University 1%
Student Paper

5 Submitted to Jacksonville University 1%
Student Paper

6 www.aclweb.org 1%
Internet Source

7 Submitted to University of Hertfordshire 1%
Student Paper

8 "Combating Fake News with Computational Intelligence Techniques", Springer Science and Business Media LLC, 2022 1%
Publication

www.researchgate.net

9

Internet Source

1 %

10

Submitted to New York Institute of Technology

Student Paper

1 %

11

arxiv.org

Internet Source

1 %

12

sjcit.ac.in

Internet Source

1 %

13

Afrin Jaman Bonny, Puja Bhowmik, Md. Shihab Mahmud, Abdus Sattar. "Detecting Fake News in Benchmark English News Dataset Using Machine Learning Classifiers", 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2022

Publication

1 %

Exclude quotes Off

Exclude matches < 1%

Exclude bibliography On