Interpretable Bangla Fake News Classification Using BERT and Traditional Machine Learning Approaches

by

Ramisa Anan
19201101
Elizabeth Antora Modhu
18301075
Arjun Suter
18101419
Ifrit Jamal Sneha
19201136

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
September 2022

Declaration

It is hereby declared that

1. The thesis submitted is my/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:

Ramisa Anan	Elizabeth Antora Modhu
19201101	18301075
Arjun Suter	Ifrit Jamal Sneha
18101419	19201136

Approval

The thesis titled "Interpretable Bangla Fake News Classification Using BERT and Traditional Machine Learning Approaches" submitted by

1. Ramisa Anan (19201101)
2. Elizabeth Antora Modhu (18301075)
3. Arjun Suter (18101419)
4. Ifrit Jamal Sneha (19201136)

Of Summer, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on September 20, 2022.

Examining Committee:

Supervisor:
(Member)

Annajiat Alim Rasel
Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Co-supervisor One:
(Member)

Dr. Matin Saad Abdullah
Professor
Department of Computer Science and Engineering
BRAC University

Co-supervisor Two:
(Member)

Mr. Moin Mostakim
Lecturer
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, Ph.D.
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

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

Ethics Statement

Hereby, we ensure that the following is true for the manuscript "Interpretable Bangla Fake News Classification Using BERT And Traditional Machine Learning Approaches":

1. This content was created entirely by the authors and was never before published.
2. There are no plans to publish the work elsewhere.
3. The author's own research and analysis are fully and accurately reflected in the work.
4. Co-authors and co-researcher's key contributions are carefully acknowledged throughout the paper.
5. The findings are correctly positioned in relation to earlier and ongoing studies.
6. All sources are gratefully acknowledged. If a text is copied verbatim, it must be acknowledged as such by using quotation marks and providing the appropriate citation.
7. All authors will accept public responsibility for the paper's content because they were all personally and actively involved in the substantial work that went into it.

There could be serious repercussions if the ethical statement standards are broken.

Abstract

Fake news is a type of content that is inaccurate or misleading and it is usually published with the intention of damaging a person or organization's reputation. It has recently grown significantly in the online forum and on social media platform like Facebook, Reddit, Twitter etc. Because of its falsified statements, people are often persuaded by false news, which has serious consequences in the real world. As a result, there is a growing interest in the field of fake news identification, even though the majority of fake news identification studies are for English language whereas just few of them are for Bangla language. In our study, we come up with a BERT-based system that uses Stratified K-fold cross validation that can achieve 98.45% test accuracy, whereas only the Random Forest can achieve 86.83% accuracy among all the traditional machine learning models. Furthermore, we used Local Interpretable Model-Agnostic Explanations to provide explainability to our system. In this research, we have used the existing BanFakeNews dataset to identify Bangla Fake News. The primary focus of this paper is to develop a model that can recognize fake news in natural language processing so that the developed model can decrease the time it takes individuals to extract fake news from social media.

Keywords: Bangla Fake News, Natural Language Processing, BNLP, Traditional Machine Learning, BERT.

Acknowledgement

Undoubtedly, interdependence is better than independence. The true spirit of accomplishing a goal is through following a path of excellence and ongoing discipline. A journey is always better when traveled with dynamism. Without the assistance and inspiration from numerous individuals, we would never have been able to finish our task.

First and first, we thank the great Almighty, without whose grace our thesis could not finish without significant setbacks. We render our utmost gratitude to the almighty for whatever we have achieved.

Second, we thank our worthy and respected supervisor, Mr. Annajiat Alim Rasel sir, Senior Lecturer, Department Of Computer Science and Engineering for his thoughtful and constant guidance during our research. He came to our help anytime we needed it. We also would like to thank our co-supervisors Dr. Matin Saad Abdullah, Professor and Mr. Moin Mostakim, Lecturer for their counsel.

Lastly, we are appreciative of our parents. Without our parents' ongoing support, it might not be achievable so thanks to their kind prayers and support.

Table of Contents

List of Figures

List of Tables

Nomenclature

The list describes several abbreviation that will be later used within the body of the document.

AP Average Precision

AUC Area Under The Curve

$BERT$ Bidirectional Encoder Representations from Transformers

DT Decision Tree

KNN K-Nearest Neighbor

$LIME$ Local Interpretable Model-Agnostic Explanations

LR Logistic Regression

MNB Multinomial Naive Bayes

NLP Natural Language Processing

PR Precision Recall

RF Random Forest

ROC Receiver Operating Characteristic

SGD Stochastic Gradient Descent

SVM Support-Vector Machine

Chapter 1

Introduction

1.1 Introduction

Fake news is one kind of information that has been purposefully fabricated to misinform or manipulate the readers. Fake news contains stories which are written with the intention to mislead people and promoting a specific agenda's skewed point of view. Fake news detection is the process of identifying news that is intentionally spread through news media or social media platforms online. The primary goal of fake news detection is to assist users online in avoiding various types of fake news. Over the past several years, social media and internet news outlets have gained prominence in Bangladesh. Moreover, it is relatively simple to access both the news portals and the social media pages that are affiliated with them. Both of these platforms share all elements of news from all around the nation.

The news portals are a part of online communication medium for all type of internet users. The establishment of a news portal allows for the publication of press releases, publications, articles, blogs and other news-related information. Online news portals now play a major role in informing, circulating and teaching the general audience about current events throughout the whole world [1]. Additionally, people do not have enough time to catch up on what happened the day before by reading the newspaper. Therefore, they depend on web portals or electronic media to stay updated with current events. Any story, article, or data released on a news website or social media platform has a profound impact on the general public [2]. The information we consume influences our ability to make decisions, and our worldview [3]. This is why fake news can be a major threat. Online news portals have made our lives easier as we can get access to the daily news from anywhere we want. Despite the advantages, the biggest flaw of online news portals is fake news.

Fake news is created for commercial or other reasons in order to make money or manipulate people's minds to make them believe a specific viewpoint [4]. Also, fake news isn't always just lies; it's more frequently a mix of lies and reality. This incorrect and inaccurate information is intended to deceive readers. This type of news is all over social media these days. Fake news on social media, can spread like wildfire in a matter of hours, has the potential to devastate our society as well as the whole country [5]. Additionally, it is demonstrated in a study that people tend to spread fake news more frequently than real news and this is another major downside of spreading fake news [2].

The news portal websites are open to everyone and accessible from any device with

internet connectivity. The ease of access to information is not an issue, but it can be problematic if it contains misleading information or reports that spread throughout the world. Fake news can damage one's personal and social life, as well as contribute to political chaos and misinterpretation. In Bangladesh, circulating fake news is a pretty common phenomenon. The outcome of fake news is always upsetting and resulting in massive loss. Lynching and violence have become major concerns in every country as a result of the spread of false information. This is why detecting fake news has become a crucial challenge. Furthermore, developing a model which can detect Bangla fake news is necessary to stop the dissemination of misinformation in our country.

Although there are many advanced models for detecting fake news, those models mostly emphasize the English or other languages. In Bangla language, there aren't many models that can reliably detect fake news. Therefore, our primary goal for this study is to develop an effective model which can recognize fake news in Bangla language. We used traditional machine learning algorithms, BERT with Stratified K-Fold to create comparison of detecting fake news.

The key goals of our study are as follows: (i) Develop an automated fake news detection system in Bangla that can be used in various NLP-based systems such as text-based news classifiers. (ii) Improve the overall efficiency of the process by doing research to determine the best framework. While doing the study, we discovered a drawback that the most prior research for detecting false news have been conducted in English, thus we sought to focus on Bangla language.

The contribution of this article are as follows:

- A comprehensive study of several machine learning models for detecting fake news using the BanFakeNews dataset.
- BERT with stratified K-fold cross validation model has been developed with the purpose of detecting fake news.
- To investigate the interpretability of the proposed model, Local Interpretable Model-Agnostic Explanations are being used.

In this paper, chapter 2 discusses related work in the domains of fake news detection. The models used to train the BanFakeNews dataset, as well as an instance of the models are described in chapter 3. The 4th chapter discusses the findings and analyses. Chapter 5 concludes with a hope for enhanced model performance and future work in the same field.

1.2 Aims and Objectives

We see how often we get fake news from a Facebook group or the online news portals. These news get circulated around very quickly. Since these fake news seek to propagate false information in news content and mean to harm social peace and harmony, we aim to conduct a comprehensive study of machine learning models to identify Bangla fake news.

All through our survey, we saw that a large portion of the research works present a dataset proper for the methodology they are using and some available dataset is there just based on unambiguous review points. In this paper, we have addressed

the issue of fraudulent news detection, particularly in Bangla language using news portal dataset. The focus of our research is to apply multiple types of deep learning as well as machine learning approaches to identify if there is fake news and explore the results of these approaches. Therefore, our research objectives are-

- To filter the news whether it is an authentic one or fake
- To minimize the error and noisy data
- To provide suggestions for enhancing the proposed models
- To assess the proposed models for further improvements
- To find a model that performs relatively better

We expect this work will assume an indispensable part in the improvement of phony news identification frameworks.

1.3 Problem Statement

Fake news is certainly not a modern phenomenon. Fake news has emerged long before the printing press was established. As long as people have lived in social groups where power dynamics are important, rumors and misleading information have most likely existed [6]. News was typically spread from person to person orally before the printing press was created. The reliance on social media for news and information has both advantages and disadvantages. However, people use social media to find and perceive news because of its cheap cost, simplicity of access, and rapid dissemination of information. On the reverse side, it enables the propagation of false news with intentionally erroneous material [7]. The breadth and depth of misleading information have the ability to be extremely harmful to individuals and the community. As a response, monitoring of phony news posted on social media has recently arisen as a widely discussed topic of research [8].
Because of the unique qualities and problems that make automated detection difficult, present classification method using traditional news sources are useless or unsuitable for recognizing phony data from online [8]. First of all, fake news's are not only challenging but also time-consuming to recognize depending on media information since it is purposefully produced to encourage readers to believe incorrect information. Therefore, in order to make a decision, we also need to take into account auxiliary information, like user social interactions on social media. Moreover, exploiting this accessory data is problematic to use since consumers' social interactions with fake news gather information which is extensive, fragmented, unpredictable, and confusing [9].
Although spreading false information is one of the inaccurate ways of damaging the reputation of someone or any entity, these recently opened numerous news portals dare to post false articles just to gather more audience interaction [2]. On the other hand, audiences on the internet fail to understand and fall for such misleading content because of not being able to authenticate them. The editorial team is supposed to be more cautious before deliberately promoting disinformation. An article published in Forbes makes readers follow a simple technique "Think before you click" [10].

The majority of readers are now technologically savvy, which is one of the reasons why every major and small media organization has gone digital. Despite the initiatives of prominent Internet news platforms to boost public trust, there remains widespread worry about disinformation and deception. Many people just cannot seem to reply on the news portals. According to Digital News Report 2019 survey by Reuters Institute, In Brazil, 85% of people share the view that they are concerned about online fake news. Following the US (67%) and the UK, which both have significant levels of concern, then comes Germany (38%) and the Netherlands (31%) [11].

In Bangladesh, the Bangla news portals have become quite popular in last few years. As the sites are easily accessible, spreading of fake news is very common now-a-days [12]. For example, in 2020, a well-known news portal page 'BD FactCheck' spread fake news about common people not getting COVID-19 vaccine which created unrest situation among people [13]. Moreover, in 2019, a fake report about Padma Bridge authorities risking human life on the construction site went viral on the internet. As a result of this information, random people were suspected of being child kidnappers and even beaten to death [14]. Therefore, identifying fake news in Bangladesh has become one of the toughest challenges faced in the internet.

Chapter 2

Related Work

2.1 Literature Review

Initially, researchers believed that bots may be accountable for spreading misleading content faster, so they utilized advanced bot-identification technology to eliminate web-based entertainment shares produced by bots. Whereas, the result seemed to be quite the same, False news spread generally as fast as before. Which implied that individuals were answerable for the share-ability of misleading news. There have been very few drives taken to accomplish fake news recognition. Recognizing the effect of phony news, specialists are attempting various procedures to track down a fast and programmed answer for identifying counterfeit news in the recent few years. M. G. Hussain et al. in their paper tries to analyze Bangla fake news as there is not much work that has been done on it previously. As of the shortage of data, they stripped articles from renowned news portals such as- Prothom alo, Kaler kantho etc. And collected almost 2500 news article datasets. In their paper they used SVM and MNB classification models sequentially and observed that SVM achieved a 96.64% accuracy rate which is better than MNB model with the accuracy of 93.32% [15].
R. R. Mandical et al. in their report develops a system that can consistently categorize bogus news to enhance the process of characterizing it. They aimed to extract fake news through Deep Neural Networks, Passive Aggressive Classifiers, and Naive Bayes classifiers. While using seven datasets that are being gained from an assorted arrangement of sources and found that some particular datasets have performed fundamentally better compared to others. They also observed that the DNN model surpassed both naive Bayes and passive-aggressive classifiers in all but one sample [16].
In the paper, H. Bingol et al. extracted data from social media for the classifier. They implemented Naive Bayes, Random Forest, SMO, RL , OneR, JRip, and ZeroR for solving the rumor detection task. Among them, SMO and Random Forest got better results with an accuracy of 98.7% [17].
Additionally, I. Ahmed et al. have used articles from the internet as classifiers by using ML ensemble techniques. The research has explored four different real-world articles. Where three of the datasets are available on the internet with a number of 44,898 and 20,386, and 3,352 articles respectively and merging those three datasets they got their fourth one. Their proposed combined model performed better than any individual models while achieving 99% accuracy by Perez-SVM and random forest algorithm [18].

On the other hand, A. A. Imran et al. applied a study on both the real and fake Bangla news available on the online news portals. Among seven significant ML algorithms, DNN performed the best with 90% accuracy [19].
In a study, F. Harrag et al. proposed another model which can automatically identify if the given Arabic news or claim is authentic via a deep neural network using CNN. Their methodology tries to overcome the problem through fact-checking. Surprisingly, this model exceeds the performance utilizing the aforementioned dataset of the state-of-the-art method with almost 91% of accuracy [20].
However, Twitter is no less than other platforms for posting misleading content and using tweets and comments. Research shows that tweets that include false claims reach six times quicker to its users than honest tweets [21]. In 2013, there was a rumor tweet about an explosion that happened at the White House which caused President Barack Obama an injury to the Associated Press (AP) Twitter account. It was claimed to be false information after causing huge instability in the stock market [22]. In a paper, A. A. Tanvir et al. extract twitter posts as their datasets to predict fabricated news messages. This study executes among five notable ML algorithms separately to illustrate the productivity of the classification problems on the same datasets where they found that Naïve Bayes and SVM achieve better results than other algorithms with F1-score 0.94 both [5].
J. A. Nasir et al. proposed a hybrid deep learning model which combines CNN plus RNN archives for better results while detecting fake news in the English language. This research contains d1(FA-KES) - 804, d2(ISOT)- 45000 news articles as datasets. Eventually, this dataset had gone through further data splitting and pre-processing techniques for model implementation. This study found the proposed Hybrid CNN and RNN model achieved accuracy of 99% (on ISOT dataset) which is better than any other supervised classifier method [23].
Identifying fake news from English tweets is a functioning active area of research. Also, a number of research has already been done in this field. However, from the linguistic perspective, the ratio of detecting fake news in the Bangla Language is much lower than expected. Verifying fake news in Bangla is quite a tough task because of the less availability of the dataset. In particular,to identify phony news M. Z. Hossain et al. created a standard system written in Bangla particularly. Also, they brought an in-depth survey of the result they got with human performance regarding the detection of misinformation. For their research, they gathered 50,000 news in their dataset. They have achieved best results while incorporating all linguistic features with SVM that shows a f1 score of 0.91 [24].
In the paper, E. Hossain et al. used 57000 Bangla news dataset and applied K-fold cross-validation on Bi-LSTM with the Glove along FastText model searching for Bangla fake news. This study got 95% and 94% accuracy rates concurrently through training, they also used GRU in their study which shows 77% accuracy [25].
Moreover, F. Islam et al. analyze consideration of the South Asian context, they explore Bengali false news categorization. 726 news articles were retrieved from facebook that are tagged as fake for their datasets. This paper has used Random Forest and Logistic Regression which demonstrates a decent response of 85% and 77% accuracy, accordingly [26].
In another paper, T. Islam et al. observed Facebook and Youtube comment for scam filtering with text documents within the fake bangla context. They collect 1965

public data by extracting comments on Facebook and Youtube for the experiment. And applying Multinomial Naïve Bayes (MNB) they got an accuracy of 82.44% [27]. M. Z. H. George et al. used a dataset containing around 50,000 Bangla news and proposed a hybrid model of CNN-LSTM architecture for Bangla fake news or data identification. The model obtains 75.05% accuracy by initializing its work by collecting data through websites while using a deep learning methodology [28].
However when the World Health Organization (WHO) proclaimed COVID-19 an outbreak, people started to assume negative situations and start posting without even justifying its roots or the truthfulness of their shared posts. A massive amount of false news has been spread in Bangladesh via social media, as well. Considering that P. B. Pranto et al. in their research paper, worked with such fake news that is spread over the internet and social media platforms and developed a model to identify Bangla incorrect facts during COVID-19. This study has a collection of 3187 posts in total for their datasets. They used three models for the experiments: BERT, XLM-RoBERTa, and DistilBERT. Eventually found BERT is the top performing model, with an F1-score of 0.97 . Besides, XLM-RoBERTa archives a F1-score of 0.95. On the other hand, DistilBERT gets a F1- score 0.91 [29].

2.2 Related Work Summary Table

Model	Datasets	Accuracy	Reference
Support Vector Machine	Bangla News Articles	96.64%	[15]
Deep Neural Network	Online News Portals	90%	[19]
Sequential Minimal Optimization and Random Forest	News Articles	98%	[17]
Random Forest and Perez-LSVM	ISOT	99%	[18]
Naive Bayes	Twitter Posts	94%	[5]
Hybrid Convolutional Neural Network + Recurrent Neural Network	News Articles	99%	[23]
All linguistics features with Support Vector Machine	News Articles (BanFakeNews)	91%	[24]
Bidirectional-LSTM	Bangla News Articles	94%	[25]
Random Forest	Fake News Articles from Facebook	85%	[26]
Multinomial Naive Bayes	Public Comments from Facebook and Youtube	82.44%	[27]
Hybrid Convolutional Neural Network + LSTM	Bangla News	75.05%	[28]
Bidirectional Encoder Representations from Transformers (BERT)	Comments form Social Media Platforms	97%	[29]
Deep Neural Network + Convolutional Neural Network	Arabic News	91%	[20]

Table 2.1: Comparison of Various Studies on Fake News Detection

Chapter 3

System Model

In this chapter, We get an insight of our proposed Fake News detection model, which is divided into three parts. This paper's section 3.1 outlines our proposed model. On the other hand, section 3.2 is broken into two subsections. The first subsection discusses the data description, while the second subsection discusses data preprocessing. Finally, section 3.3 reveals our model's specifications.

3.1 Proposed Model

Initially, we choose to use the BanFakeNews dataset to examine our model. While working with text in Natural Language Processing, text cleaning is a very essential step. As it is an existing dataset, we had to consider a range of preprocessing procedures to clean the data because the obtained data can be often noisy. Next, we have divided the data into two portions to use the traditional split and stratified K-Fold, where the stratified K-Fold is employed with BERT in order to feed the data to our preferred model. Moreover, we have the typical split of three parts: training, testing, and validation, where the training set includes 60% of the data, while the test and validation sets each include 20%. On the labels for each of the three split sets, we then ran one hot encoding. One hot encoding is a method of preparing data for an algorithm as well as it improves prediction. The train and test sets were later fed into machine learning models, including Random Forest, Logistic Regression, Decision Tree, K-Nearest Neighbor, Stochastic Gradient Descent, Support Vector Machine, Multinomial Naive Bayes as well as stratified K-Fold as Cross-Validation with BERT. We assessed the efficiency of our trained model using the validation set after training the model. Alongside the ROC curve and confusion metrics, we have also used a variety of performance measures in order to determine which model tends to be more efficient in terms of accuracy, Recall, precision, F1 scores. We have undertaken a thorough comparison between the trained models with one another. We evaluate the model and those assessed using a test dataset to discover whichever model works better in detecting bogus news. Lastly, we export the best suited model based on our analysis. We applied LIME with the exported model, which helped in adding explainability to our analysis. Importantly, it is an approach for approximating any black box machine learning model and explaining each individual prediction using a local, interpretable model. This helps us understand how our model analyzes and categorizes a prediction owing to its explainability. Figure 3.1 shows the workflow diagram of our proposed model.

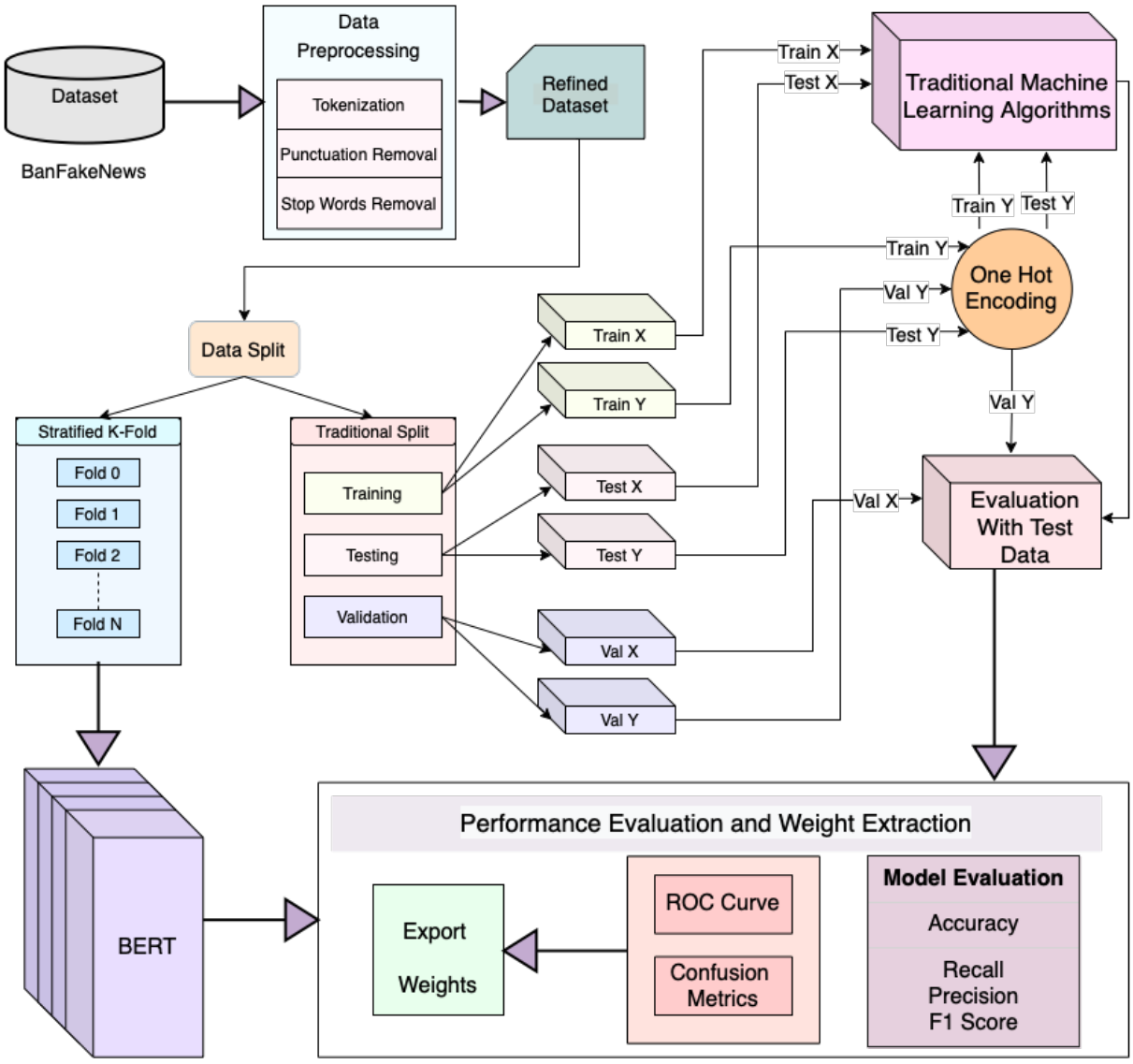

Figure 3.1: Proposed System Model: Interpretable Bangla Fake News Classification

3.2 Data

In this study, we used the BanFakeNews Dataset in our implementation model. We utilized the LabeledAuthentic-7K and LabeledFake-1K files from the dataset. Afterwards, we merge the two datasets, then analyze and process them before applying them to our suggested model.

3.2.1 Data Description

The paper extracted their data from the public and most renowned online news portals of Bangladesh. Among all the available news websites, they choose 22 most highly regarded and well reliable news platforms for gathering authentic news. On the other hand, they had three criteria for collecting fake information, since they collect their data from online sources. Such as news that includes erroneous or deceptive context, news with sarcastic comments and clickbait which are commonly used as attention grabbers for the audiences. However, renowned web portals tend to publish sarcastic news. Sarcastic comments regarding famous personalities are supposed to spread faster than usual. Similarly, clickbaits are clickable links with some interesting and provoking information to increase visitors to some particular sites. Which are more likely to be found in smaller and maybe less recognized news companies. Furthermore, they eliminated redundant data after retrieving news from all these online sites because a number of diverse sources contain a kind of similar mockery or misleading news, which raised the possibility of having the precise same information. Since, the paper has considered anything regarding false news and sarcastic news as the fake news. Therefore, they found few websites that provide a rational and detailed description of misleading information, which has previously been reported on other web pages. As a result, we decide to work with this dataset as it may be the most appropriate one for our model [24].
The dataset contains 8501 data points, with around 7000 categorized as real news and 1000 identified as fake news. Even though our dataset is not perfectly balanced, we may conclude that its quality is moderate. The dataset has been classified as '0' and '1', where '0' representing fake news and '1' representing legitimate news. The information includes all the news context, headlines, publishing times, domains etc., and it has been grouped into total 12 categories. We used the news headline of the dataset to detect false news, since news headlines convey more about news authenticity. To acquire a comprehensive picture of the data, we thoroughly analyzed the length of news headlines. The headline length of the real news is shown in Figure 3.2 and Figure 3.3 depicts the headline length of fake news.
Figure 3.4 and 3.5 are authentic and fake news corpuses. The term "Data Corpus" refers to a group of spoken or written words that can be utilized for a number of purposes, such as offering information on how language can be used. One of the benefits of using the corpus is that we can accurately estimate our model if we can determine the data value obtained by combining the two corpora. These two data corpora can be used in the machine learning models to extract information from the data and feed it back into the data corpus.
Figure 3.6 shows the word count for the training phase of the BanFakeNews dataset, which can give us a rough estimate of how many words are identified for fake and authentic news individually. Knowing the word count will help us determine the

precise quantity of data we can incorporate into our model after data preprocessing. Next, figure 3.7 provides us with the count of unique words in both authentic and fake news in the training set. Figure 3.8 shows the count of stop words in both of the corpora. Figures 3.9 and 3.10 illustrate the URL count and mean word length count in the training set for the authentic and fake news corpuses, respectively. Figures 3.11 and 3.12, respectively, indicate the character and punctuation counts for both the corpuses. Lastly, figure 3.13 and 3.14 are showing the hashtag count and mention count for both authentic and fake corpuses. These figures also provide an estimate of the amount of data that will be excluded during preprocessing so that we can effectively use the training set.

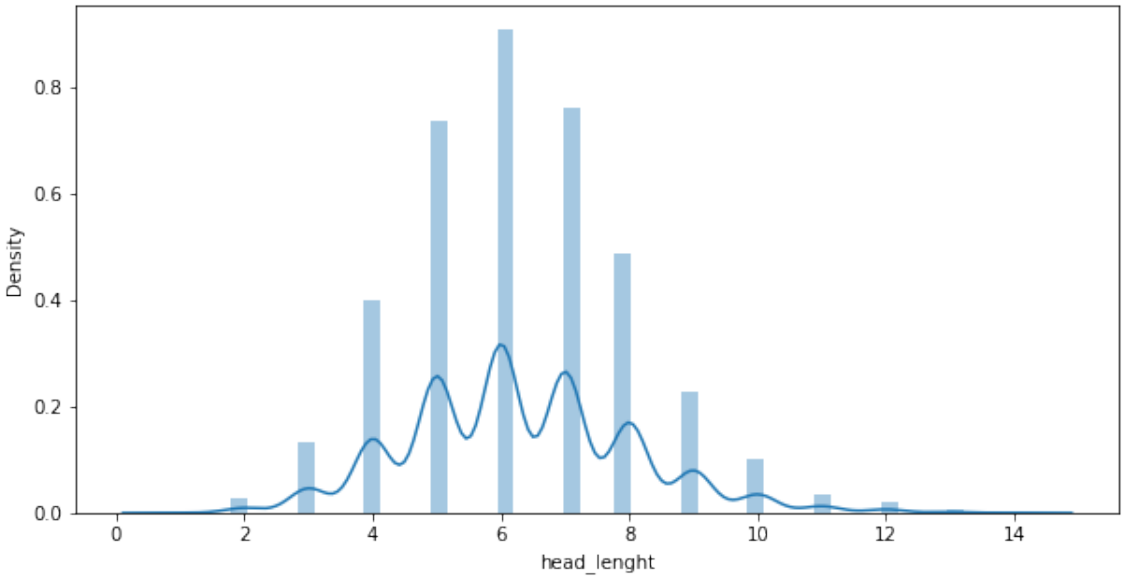

Figure 3.2: Authentic News Headline Length

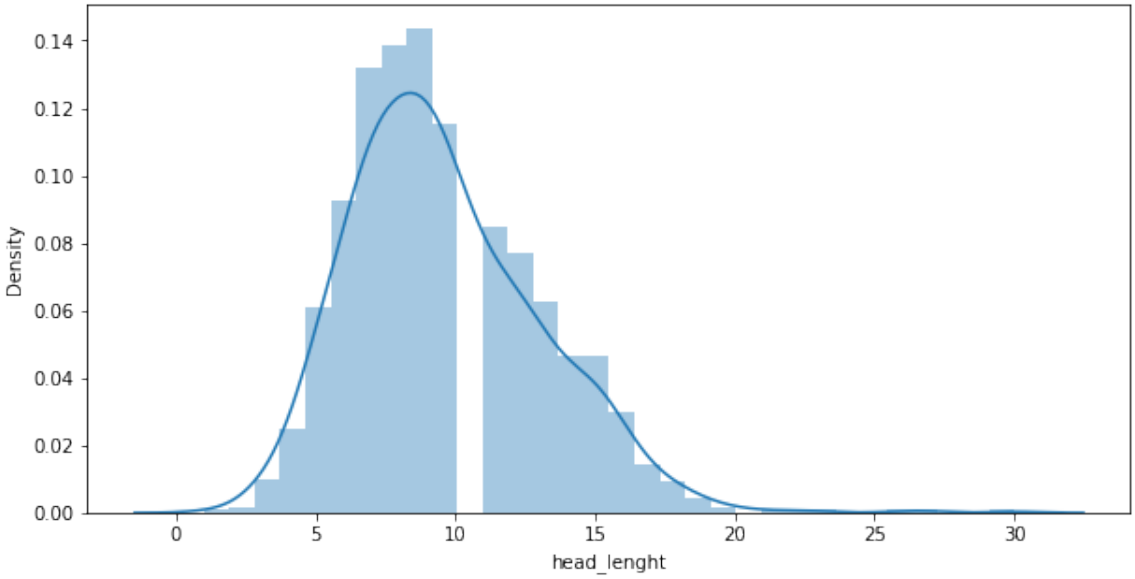

Figure 3.3: Fake News Headline Length

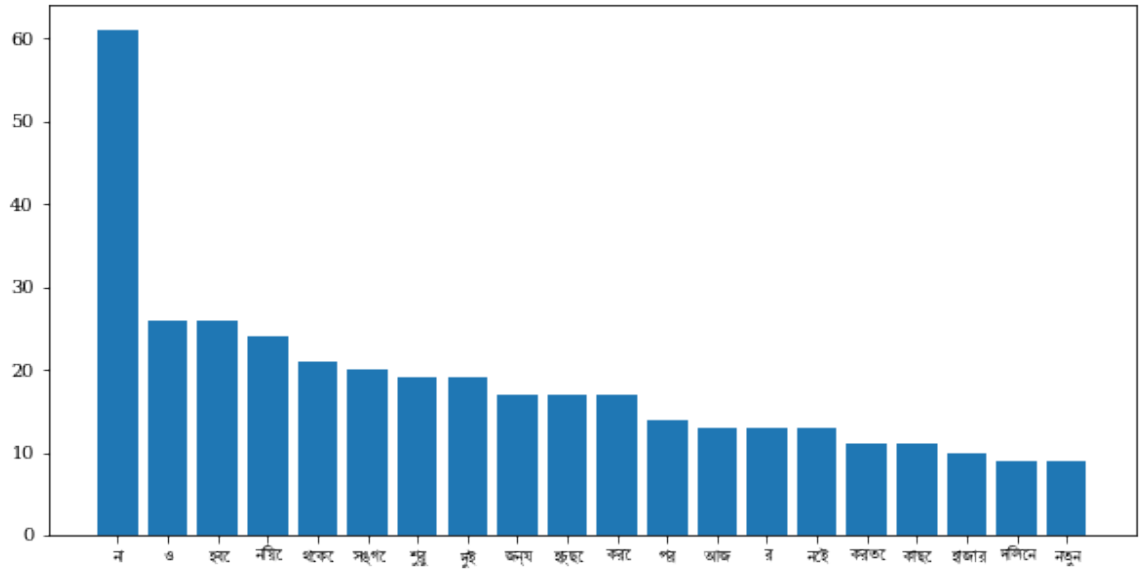

Figure 3.4: Authentic News Corpus

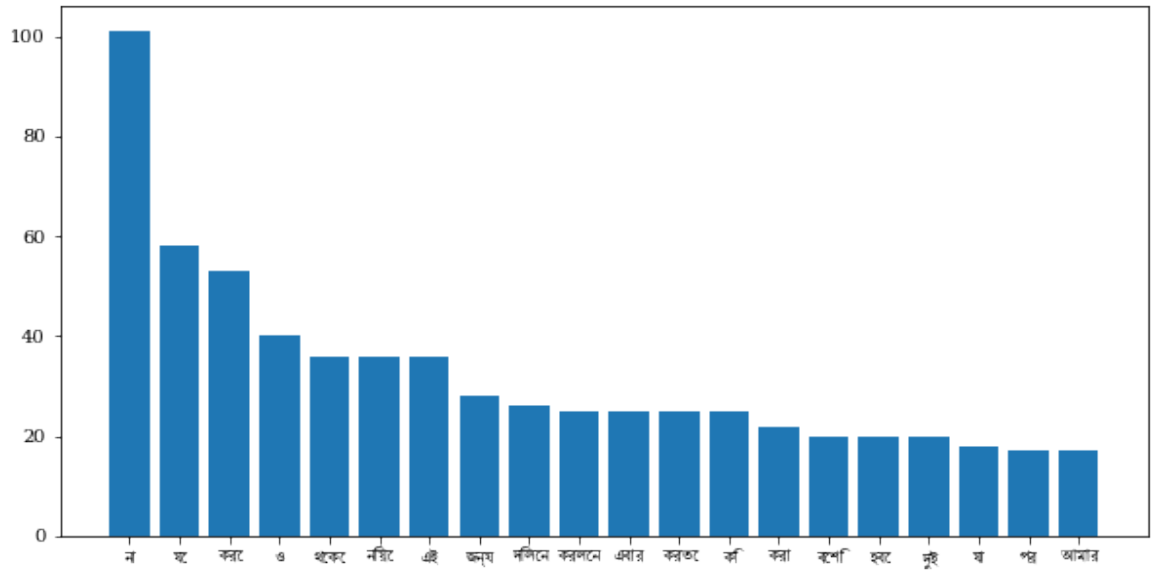

Figure 3.5: Fake News Corpus

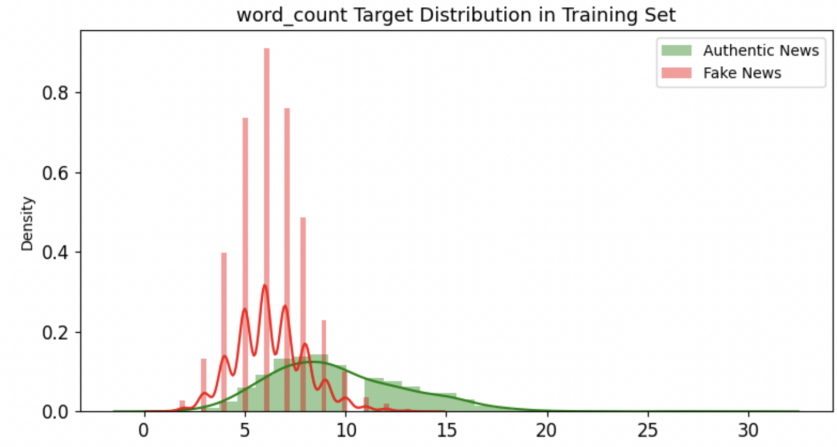

Figure 3.6: Word Count of the Dataset

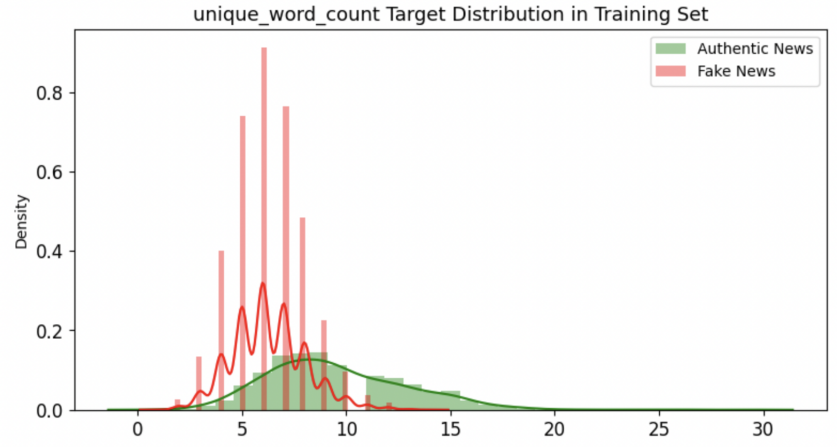

Figure 3.7: Unique Word Count of the Dataset

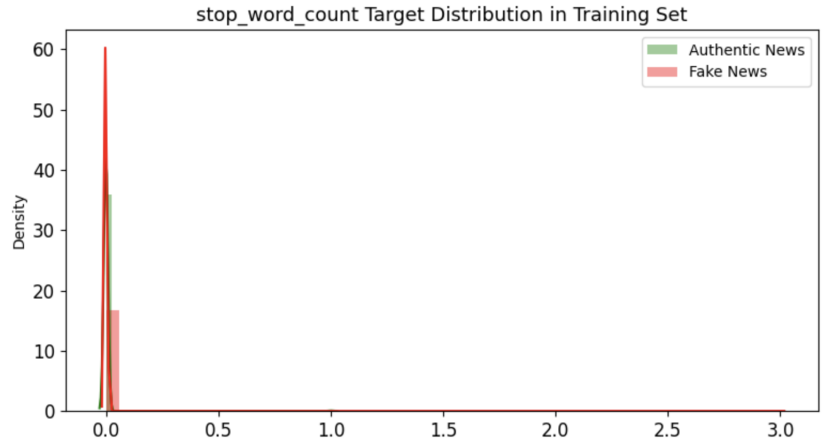

Figure 3.8: Stop Word Count of the Dataset

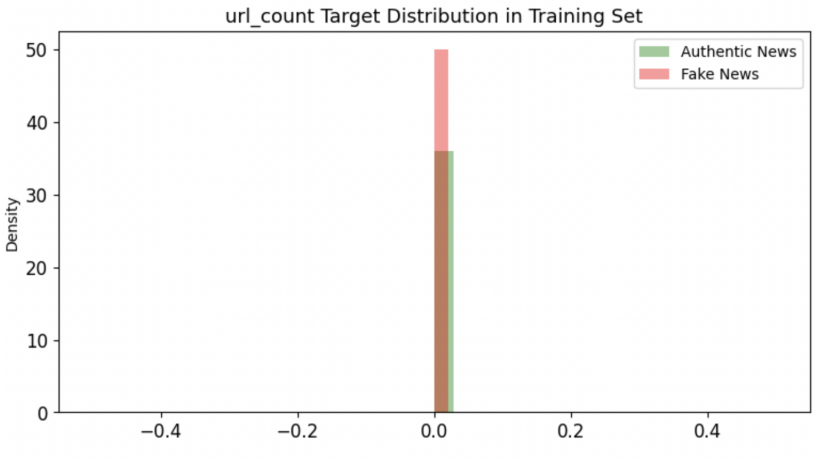

Figure 3.9: URL Count of the Dataset

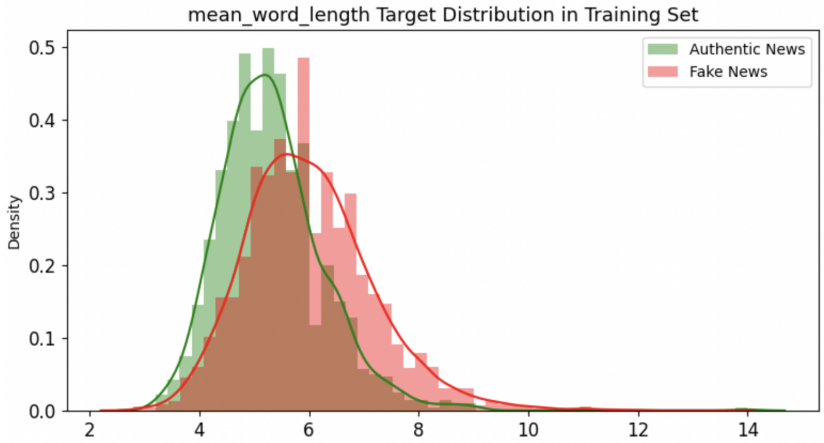

Figure 3.10: Mean Word Length of the Dataset

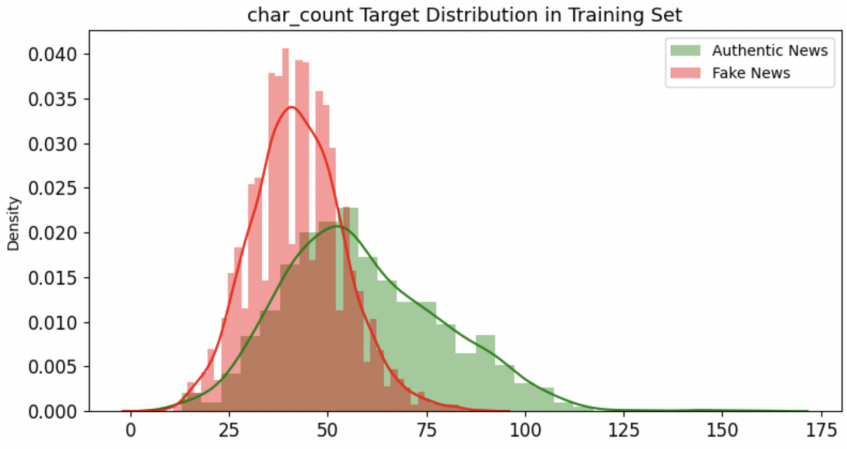

Figure 3.11: Character Count of the Dataset

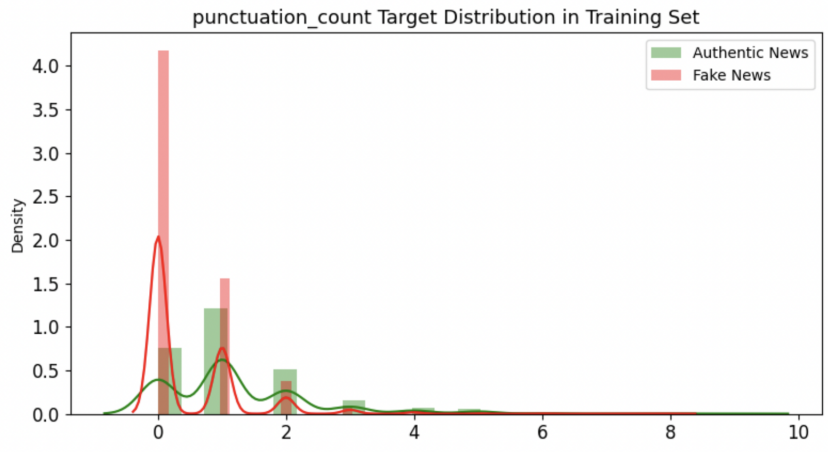

Figure 3.12: Punctuation Count of the Dataset

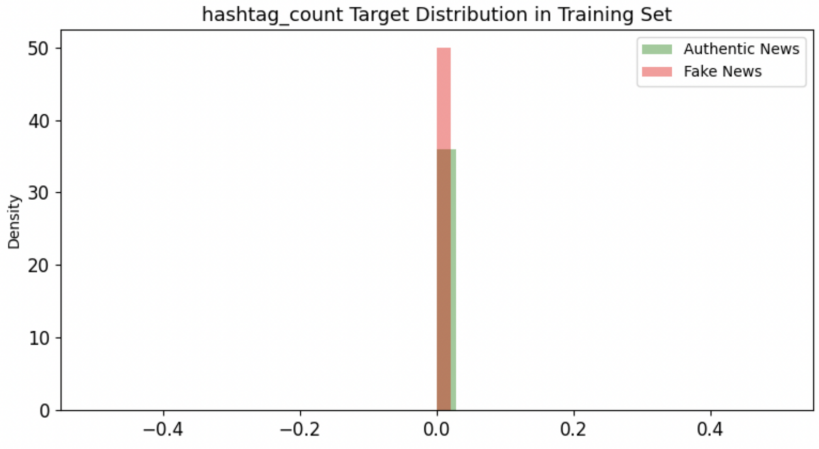

Figure 3.13: Hashtag Count of the Dataset

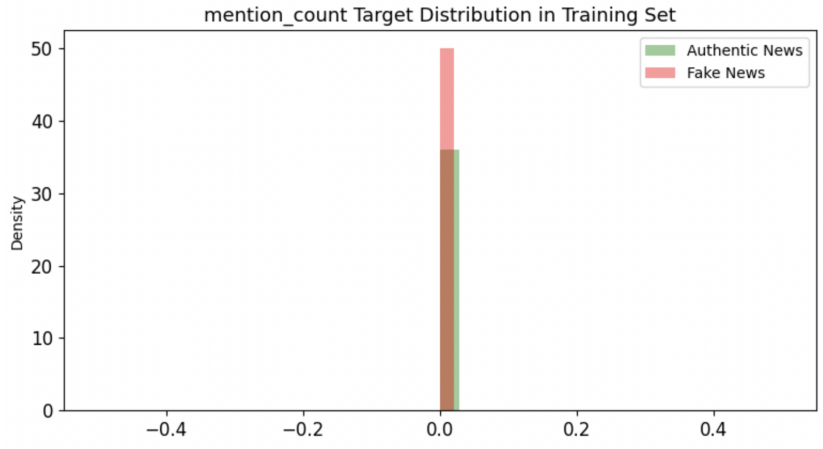

Figure 3.14: Mention Count of the Dataset

3.2.2 Data Preprocessing

Data preprocessing is a crucial stage before training any model on a dataset. Punctuation removal, stop words removal, and tokenization are the preprocessing techniques we employed. Presently, to write any word or expressions individuals mostly use keyboard character or punctuation to indicate a facial expression as an emoticon. So, the following step in data preprocessing, we remove unnecessary punctuation and tags. We eliminated those because they wouldn't provide any useful information. Stop words, which are generally very common words, are the words that are typically filtered away prior to processing a natural language dataset. In order to focus more on the vital information, we eliminated the low-level information from our data by deleting these words. There are very few tokens involved in training, that's why the removal of stop words reduces the size of the dataset as well as the training time. Tokenization splits up text into a collection of meaningful parts, or tokens. Our data is regarded as clean because we have previously removed stop words, punctuation, and emoji. After the tokenization process, we acquired character strings that are free of spaces and are considered to be tokens. We have cleaned the dataset using these data processing methods.

Before: এনআরসি নিয়ে পথে বিজেপি, পাল্টা প্রচারে তৃণমূল

After: এনআরসি পথে বিজেপি পাল্টা প্রচারে তৃণমূল

Before: তানজানিয়ায় ফেরি ডুবে ৪২ জন নিহত

After: তানজানিয়ায় ফেরি ডুবে ৪২ নিহত

Before: কয়েলের আগুনে পুড়ল রানী, মায়ের অবস্থা আশঙ্কাজনক

After: কয়েলের আগুনে পুড়ল রানী মায়ের অবস্থা আশঙ্কাজনক

Before: ষড়যন্ত্র হলে প্রধানমন্ত্রীর পক্ষে লড়বে তৌহিদী জনতা

After: ষড়যন্ত্র প্রধানমন্ত্রীর লড়বে তৌহিদী জনতা

3.3 Model Specification

Our proposed model is the traditional machine learning algorithm and BERT with k-fold validation. K-fold is a cross-validation technique used to measure a machine learning model's ability on untested data. Additionally, we employed LIME, which improves our system's interpretability. In this part we have discuss about the model specification.

Random Forest

Random Forest is a common classification method which is used in supervised learning. This method has the capability to solve issues like classification and regression in algorithms. It is based on the idea of supervised learning, which also can be used for combining several classifiers to overcome complex problems. This also helps to improve model performance. We employ this model because it takes less time to train than other approaches. This method performs greatly in scenarios where variables are much higher than observations. It may also be applied to complicated challenges and it is simple to modify for a range of ad-hoc learning activities, and gives metrics of changing relevance. The fact that RF may be employed to handle a wide range of prediction problems with only a few tuning parameters has greatly increased their appeal. The method is widely renowned for its accuracy, handling of small sample quantities, and high-dimensional feature spaces, as well as its ease of use. Because it is easily scalable, it has the ability to manage large systems [30].

Decision Tree

Decision Tree is another type of supervised learning method which is also used in both classification and regression based tasks. To put it simply, Decision Tree analysis is a divide-and-conquer technique for categorization. It is the most efficient tool for classification as well as prediction. Basically, it is a tree structure that resembles like a flowchart. Where each internal node indicates a test on an attribute and every branch expresses a test result. The leaf nodes represent a class label [31]. It has solid foundation in machine learning and artificial intelligence related studies. Decision Trees, mainly used in large databases to uncover attributes and extract patterns that are important for discriminating and predictive modeling. To these characteristics and their easy interpretation, Decision Trees have been frequently applied in exploratory data analysis as well as predictive modeling applications [32]. Furthermore, this non-parametric method can successfully handle large and complex datasets without applying any burdensome quantitative framework. If the sample data size is sufficiently large, the data can be separated into training and validation datasets. To create a viable model, a decision tree model must be built using the training dataset and the appropriate tree size must be chosen to use the validation dataset [33].

K-Nearest Neighbor

The K-Nearest Neighbors method, sometimes known as KNN, is a supervised machine learning classifier like Decision Tree. KNN has the ability to predict or categorize the way that a specific data point will be classified by proximity. All processing is deferred with KNN till the function gets assessed and the function is locally estimated. It is one of the most fundamental and basic classification techniques. For classification, this technique depends on distance. Therefore, if the features have varied weights, normalizing can greatly improve accuracy. Furthermore, if there is insufficient knowledge of the data distribution, KNN could be the initial option for classification research. The Euclidean distance in between sample set and the designated training set is frequently utilized as the foundation for the k-nearest-neighbor classifier. [34].

Multinomial Naïve Bayes

The Multinomial Naive Bayes method is mostly employed in NLP. It is a probabilistic learning technique and commonly used to address text classification issues. This Bayes theorem-based method can predict the tag of a text, like a newspaper article or an email. It also can determine the probability of the tags for a particular sample as well as return it with the highest probability [35]. For classification using discrete features, Multinomial Naïve Bayes is suitable. It is simple to implement and highly computationally efficient. The multivariate model and the multinomial model are the two event models that are most frequently employed. In general, the multinomial model—also known as multinomial naïve Bayes (abbreviated MNB)—performs better than the multivariate one [36]. Multinomial models are now thought to be the most common modeling strategy since they are more effective than multivariate Bernoulli models, which include language modeling in information retrieval [37].

Logistic Regression

Logistic regression is one of the widely used statistical techniques in research. This method is used for estimating probability of a binary output given an input variable. In the social sciences, logistic regression is frequently used to analyze results that are inherently or essentially represented by binary variables [38]. Also, it is one of the most important statistical procedures in fields including health care and pharmaceutical research, ecological studies, social statistics and economic science. Logistic regression is a part of almost all general purpose commercial statistical packages, if not all of them [39]. Binary outputs, such as either 0 or 1, positive or negative or true or false, are modeled using the most common types of logistic regression. This is another reason for which this method has gained more popularity than traditional regression (linear). Logistic regression is also effective for linear and binary classification tasks such as attack detection in cyber security. From a collection of distinct characteristics, logistic regression can also predict the likelihood that an event will occur. Furthermore, the scale of the dependent variable is 0 to 1, as the outcome is a possibility [39].

Stochastic Gradient Descent

To identify the model parameters that most closely match the anticipated and actual outputs, stochastic gradient descent is frequently applied in machine learning implementations. Although not exact, it functions well. The field of machine learning regularly uses stochastic gradient descent. It minimizes the extremely high computational cost, especially in high-dimensional optimization problems, allowing for quicker repetitions at the cost of a lower convergence rate. When training time is the bottleneck, stochastic gradient descent is used. Due to the fact that it is not necessary to keep track of which instances were evaluated in previous iterations, the stochastic technique can process examples immediately in a deployed system. Given that the examples are chosen randomly from the ground truth distribution, stochastic gradient descent reduces the overall estimated risk directly in such case [40]. Additionally, for underlying optimization issues like loss function, SGD the algorithm that is most frequently utilized [41].

Support-Vector Machines

An algorithm for enhancing a specific mathematical function according to a given dataset is called an Support Vector Machine. Generally it is applied for regression inspection and classifying [42]. SVM learns through example for labeling entities. An SVM, can be trained to distinguish between unauthorized and authorized credit card activity by studying a large number of reports of both. As an alternative, a massive database of scanned images of handwritten ones, zeros and other numbers can be used to train an SVM to recognize handwritten digits. SVM has gained popularity as a classification technique because of its excellent adaptability among a wide range of data science approaches, as well as the study of brain illnesses. Four fundamental ideas are all that are essential in order to fully comprehend SVM classification: the soft margin, the maximum-margin hyperplane, the separating hyperplane, and the kernel function [42]. The SVM's strength is in its capacity to learn data classification patterns with a balance between accuracy and reproducibility [43].

BERT

BERT is a technique for pre-training deep bidirectional representations from unlabeled text. It intends to train in all levels on both left and right context at the same time. As a result, the pre-trained BERT model may be upgraded with just one extra output layer to develop cutting-edge models. Without requiring significant task-specific modifications, such models might be utilized for a wide range of purposes, including language analysis and questionnaire surveys. BERT is both simple theoretically and effective experimentally. The application of Transformer's bidirectional training, a known concentration model, to language modeling is the fundamental technological advancement of BERT. The unified architecture of BERT across all jobs is one of its distinguishing characteristics. The final downstream design barely differs from the pre-trained architecture [44]. BERT has generated controversy in the machine learning community by showcasing state-of-the-art outcomes in a range of NLP tasks as well.

Local Interpretable Model-Agnostic Explanations

The concept of Local interpretable model-agnostic explanations is the abbreviation for LIME. In order to respond to each distinct prediction, it emulates any black box machine learning model and uses a local, explicable model. Users must be able to comprehend models for AI systems to be trusted by people. AI interpretability provides insight into these systems' internal workings and helps identify potential issues including causality, information leakage, model bias, and robustness [45]. LIME offers a broad framework for understanding black boxes and clarifies the reasoning behind predictions or recommendations given by AI. For an e-mail classification system, for instance, LIME creates a list of words from an e-mail to describe why it belongs under a certain category. It also locally approximates the classifier using an interpretable model (such as decision trees or sparse linear models) for generating the explanation.

Chapter 4

Performance Evaluation And Analysis

4.1 Performance Parameters

Accuracy

The difference between the measured value and the actual value can be defined as accuracy. In other words, the degree to which measurements or predictions closely resemble a given value is known as a system's accuracy. When true positive and true negative values are more significant than false positives and false negatives, accuracy can be a key performance indicator. Equation 4.1 illustrates how to evaluate a system's performance [46].

$$Accuracy = \frac{Correct\,Predictions}{All\,Predictions} \tag{4.1}$$

Precision

Despite the fact that precision and accuracy are closely related metrics, precision refers to how much information a value provides. As per equation 4.2, true positives and false positives can be distinguished with precision. To determine whether a particular classification model is effective, precision could be considered [47].

$$Precision = \frac{True, Positive}{True, Positive + False, Positive} \tag{4.2}$$

Recall

Recall, as specified by equation 4.3, is the percentage of correct predictions out of all the correct predictions that could've been generated by a classification method. Recall of a machine learning model is similar to accuracy and precision in that it depends on positive values while being independent of negative values [48].

$$Recall = \frac{True, Positive}{Total, Actual, Positive} \tag{4.3}$$

F1 score

A classification model's F1 score shows a reasonable balance between precision and accuracy. According to equation 4.4, the F1 score is calculated using a system's accuracy and recall value. When precision and recall values are ideal, the maximum F1 score of 1 can be obtained. Contrary to accuracy, F1 score can be a crucial performance measure when false positive and false negative values are more relevant than true positive and true negative ones [48].

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.4}$$

ROC Curve

The responsiveness and precision of the results can be affected by selecting an appropriate cut-off using the ROC curve. The capacity of a test to distinguish between the presence or absence of a particular condition is measured globally by the area under the ROC curve (AUC). The choice of a test threshold relies on the test's objectives and is often not made by weighing sensitivity and specificity equally in an effort to increase accuracy [49].

PR Curve Analysis

The precision-recall curve demonstrates the balance between precision and recall for different thresholds. High precision and low false positive rates are connected. However, a low false negative rate is associated with high recall. High recall and high precision are both indicated by a high area under the curve. When correlation to the training labels, the majority of the projected labels from a system with high recall but low precision returns many outcomes. A system with low recall but high precision, on the other hand, generates very few results, but the majority of its projected labels match the training labels. An ideal system that has excellent precision and recall will produce a lot of results. Then, these results will be appropriately classified. Precision-recall is an useful measure of prediction success when the classes are highly imbalanced. In information retrieval, recall measures the quantity of actually relevant results returned, whereas precision measures the relevancy of the results [50].

Confusion Metrics

The performance of any machine learning classification can be evaluated using a confusion matrix. It summarizes the count of accurate and inaccurate predictions made by the classifier, could be displayed in a tabular format. By calculating performance indicators like accuracy, precision, recall, and F1-score, it assesses a classification model's efficiency [51].

4.2 Traditional Machine Learning Models

On our processed dataset, we used 7 traditional machine learning models, includes RF, SVM, LR, KNN, DT, SGD and MNB. The dataset was applied to these models, and we obtained accuracy, precision, recall, and f1-score. A comparison of traditional machine learning models is shown in table 4.1. The RF model that we ran on the dataset gave us 86.48% precision, 100% recall, and 92.75% f1-score, as can be seen from all the models stated. Then, using Radial Basis Function kernel SVM, we were able to achieve 85.71% precision, 100% recall with a f1-score of 92.30%. The precision for the LR model is 84.24%, and we also receive a recall of 100% and a f1-score of 91.45%. Next, the KNN provides accuracy of 84.97%, recall of 95.88%, and f1 score of 90.10%. We deployed the DT and obtained a recall of 86.67%, a f1-score of 87.31%, precision of 87.96%. Afterwards, the SGD gave us 84.75% precision, 68.67% recall, and a 75.87% f1-score. Lastly, we used the MNB model, which provided us with 85.65% precision and 56.66% recall with f1-score values of 68.21%.

The accuracy of traditional machine learning models is evaluated in table 4.2. According to the table, RF obtains the highest accuracy with 86.83%, while SVM reaches up to 85.95%. Furthermore, LR achieve an accuracy of up to 84.24 %. The accuracy of KNN is 82.25%. However, DT, SGD, and MNB were unable to compete with the other models. The DT achieved 78.78% accuracy, whereas the SGD achieved 63.20% accuracy. Finally, MNB offered the least with an accuracy of 55.50%. Through analyzing both tables, we can conclude that the RF(Random Forest) outperforms the other models and attain a satisfactory level of accuracy in identifying fraudulent news.

A binary classification problem assessment measure is the ROC curve. This probability curve successfully separates the real components by plotting true positives vs false positives at various threshold values. Figure 4.2 shows the ROC curve of the traditional machine learning models, where Random Forest has the highest AUC value of 0.765. However, The AUC is a measure of a classifier's capacity to distinguish between categories and is used to analyze the ROC curve. An increase in AUC leads to an improvement in the model's ability to distinguish between positive and negative classifications. Therefore, it can be determined that the Random Forest model performed better than any other model in distinguishing between the positive and negative classes. Similar to the ROC curve, the PR curve is used to assess the effectiveness of binary classification systems. In figure 4.3, we can see a precision and recall curve comparing the traditional machine learning models. The precision-recall curve is created by calculating and showing the precision versus recall for a classification model at several thresholds [52]. Here, average precision (AP) is the weighted combination of precision achieved at every threshold, with the change in recall from the last threshold as the weight. We can observe from the graph that the RF has the highest AP value with 0.938, which is higher than any other conventional model.

Model	Precision	Recall	F-1 Score
Random Forest	86.48	100.00	92.75
Support-Vector Machines	85.71	100.00	92.30
Logistic Regression	84.24	100.00	91.45
K-Nearest Neighbor	84.97	95.88	90.10
Decision Tree	87.96	86.67	87.31
Stochastic Gradient Descent	84.75	68.67	75.87
Multinomial Naive Bayes	85.65	56.66	68.21

Table 4.1: Performance Evaluation: Traditional Machine Learning Models Precision, Recall and F-1 Score

Model	Accuracy
Random Forest	86.83
Support-Vector Machines	85.95
Logistic Regression	84.24
K-Nearest Neighbor	82.25
Decision Tree	78.78
Stochastic Gradient Descent	63.20
Multinomial Naive Bayes	55.50

Table 4.2: Performance Evaluation: Traditional Machine Learning Models Accuracy

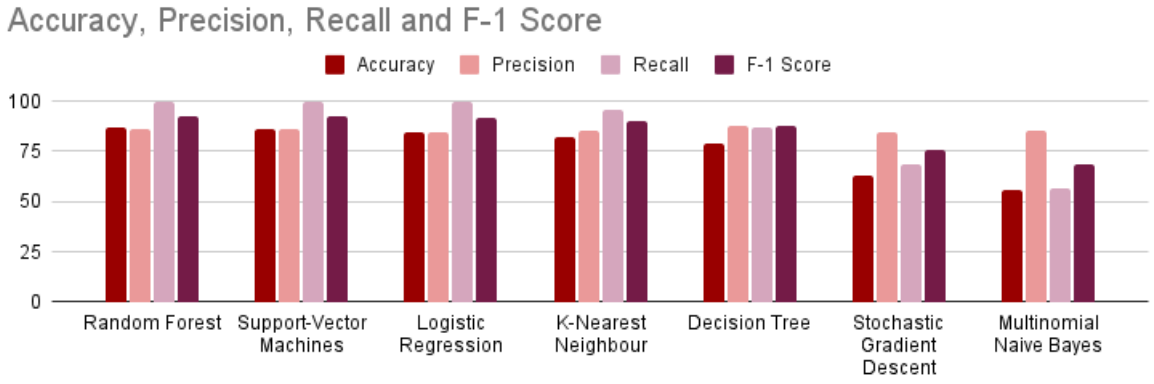

Figure 4.1: Traditional Machine Learning Algorithms Histogram

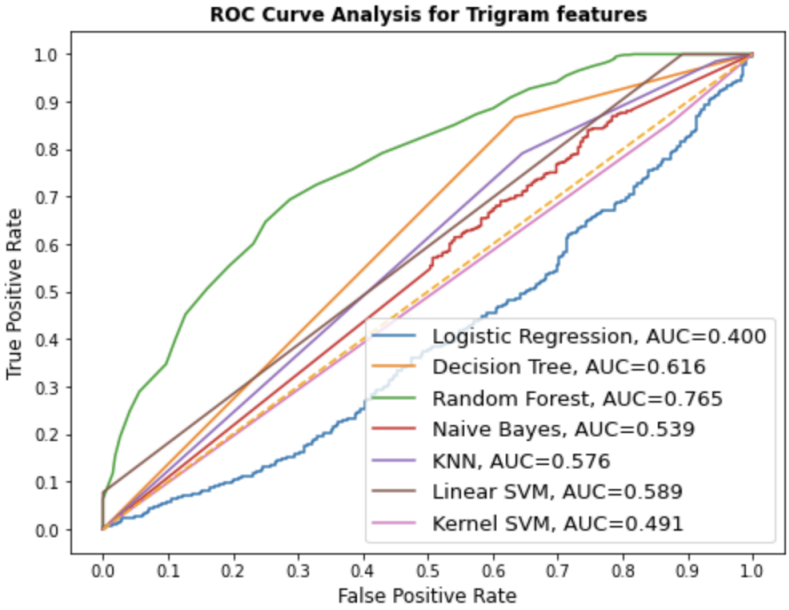

Figure 4.2: Traditional Machine Learning Algorithms: ROC Curve

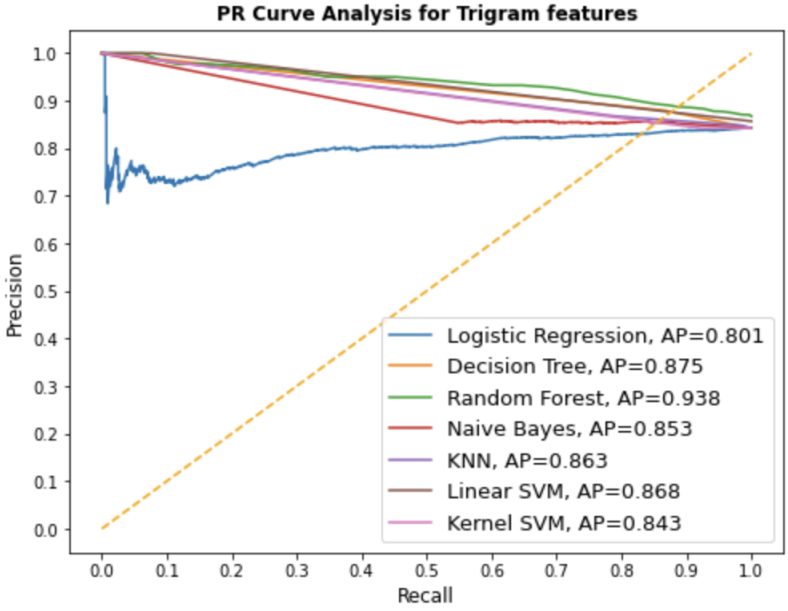

Figure 4.3: Traditional Machine Learning Algorithms: Precision Recall Curve

4.3 BERT with Stratified K-Fold

BERT is a deep learning model that stands for Bidirectional Encoder Representations from Transformers. In essence, BERT is based on Transformers. Each output unit of a transformer is attached to each input unit, and their correlation is automatically calculated based on their connection. BERT is a learning framework used for natural language processing (NLP). BERT provides context using the underlying text. So that it can aid computers in deciphering the meaning of words that are uncertain in text. The transformer is the model component that provides BERT with enhanced comprehension of verbal ambiguity and context. The transformer achieves this by examining each word in connection to every other word in a sentence, rather than processing each word separately. The Transformer offers the BERT model the ability to understand a word's full context and, as a result, better comprehend the user's intent by looking at all the nearby terms.
K-fold is a cross-validation technique for calculating how well a machine learning model performs on new data. Because it is simple to comprehend, implement, and the findings have a higher informative value than standard Classification Models, it is frequently used to evaluate a model. The data set is divided into random assign numbers in the k-fold cross validation method. When the testing set uses each fold, it splits the dataset at that step. We need to perform specific steps, such as importing the libraries required for k-fold validation on an ML model, reading and preparing the data, and implementing the k-fold cross validation method, in order to evaluate a machine learning model using k-fold validation. K-fold cross-validation helps a model by validating the data. Additionally, it makes sure that the model's performance is unrelated to the method used to select the training and test dataset. In contrast to traditional k-fold cross validation, stratified k-fold cross validation is developed primarily for classification models like BERT, in which the ratio between the training instances is the same in every fold as it is in the original dataset. We maintained the ratio between the classes in each fold while using a k fold cross valuation technique, where k is 4. We utilized the split procedure to obtain the train and test indexes for each split in order to generate the folds. Our data had to be divided into test and train data frames. To begin with, we divided the data into folds and printed the class ratios for each fold to see if they were a good representation of the entire data set. The remaining groups served as the training dataset, while each data division served as the test dataset. Following that, the BERT model is fitted to the training set and evaluated against the test set. We then kept the evaluation score from the model. We used model evaluation scores to summarize the model's performance. The table 4.3 shows performance evaluation of precision, recall, F-1 score of BERT using stratified k-fold. Furthermore, from table 4.4, we can see that, in each fold, our validation accuracy has increased significantly. In fold 0, the validation accuracy of the model was 92.80% which increased to 93.93% in the next fold. From fold 1 to 2 it became 93.93% to 95.86%. Finally, the validation accuracy increased to a maximum 98.45% in the last fold.

Fold N	Precision	Recall	F-1 Score
Fold 0	89.83	85.52	87.48
Fold 1	96.78	87.79	91.58
Fold 2	98.21	98.33	98.27
Fold 3	99.82	99.57	99.69

Table 4.3: Performance Evaluation: Precision, Recall, F-1 Score of BERT using Stratified K-Fold

Fold N	Validation Accuracy	Validation Precision	Validation Recall	Validation F-1 Score
Fold 0	92.80	88.31	82.38	84.96
Fold 1	93.93	91.72	83.63	87.02
Fold 2	95.86	91.92	92.13	92.02
Fold 3	98.45	97.40	96.56	96.97

Table 4.4: Performance Evaluation: Validation Accuracy, Precision, Recall, F-1 Score of BERT using Stratified K-Fold

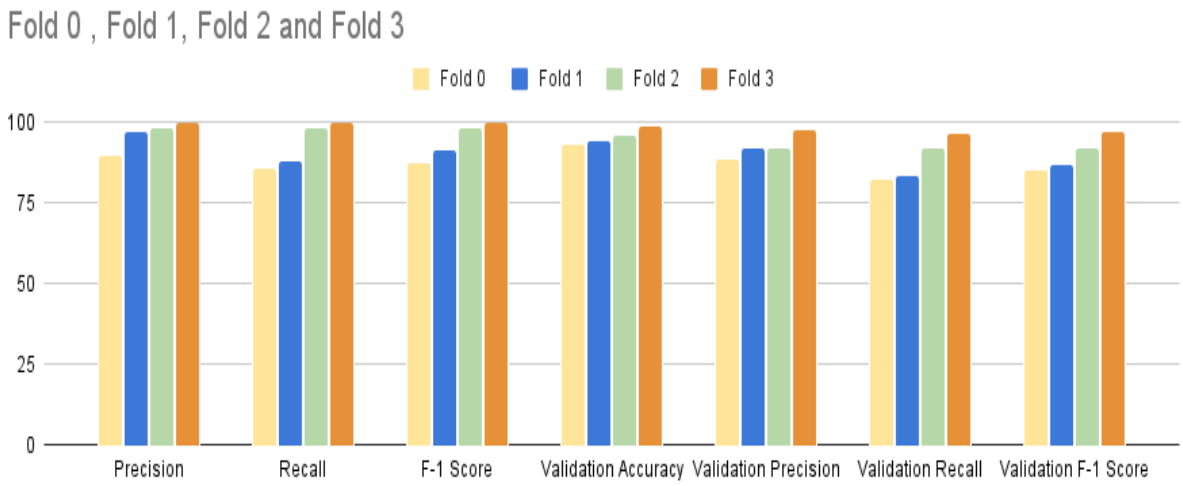

Figure 4.4: BERT with Stratified K-Fold Histogram

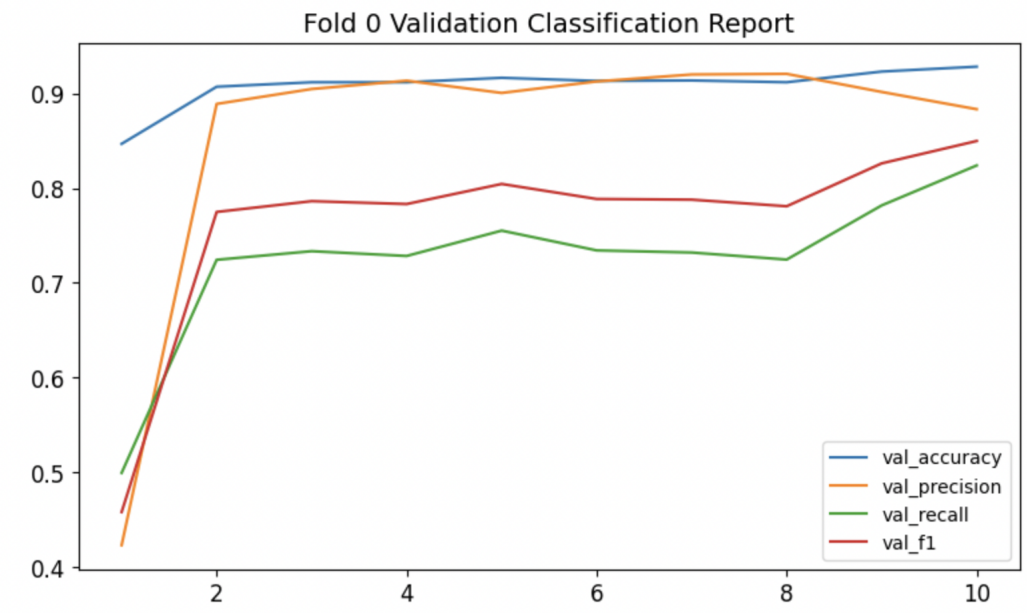

Figure 4.5: BERT using Stratified K-Fold 0: The curve of validation Accuracy, Precision, Recall and F-1 Score

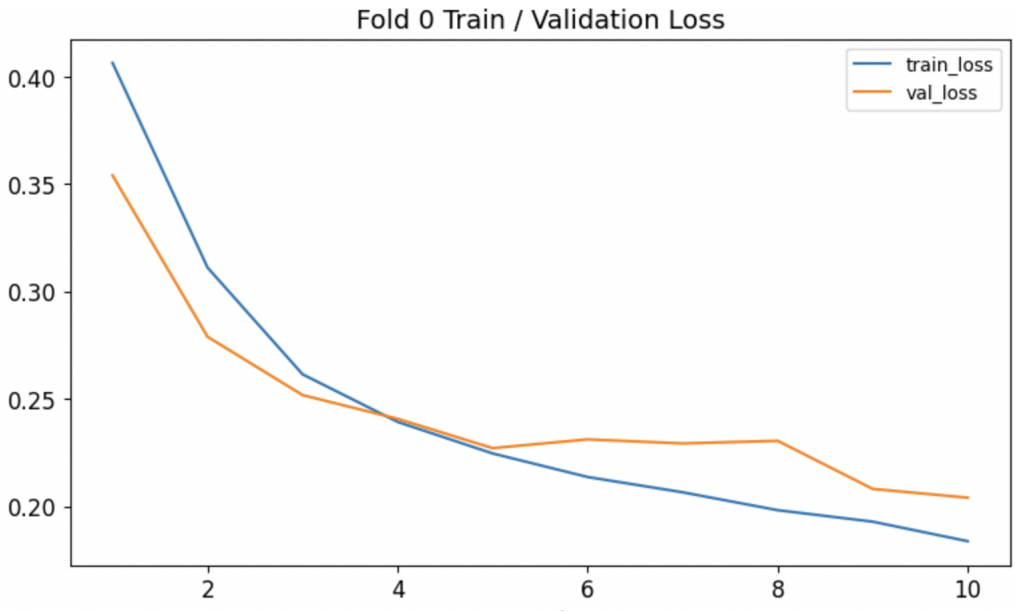

Figure 4.6: BERT using Stratified K-Fold 0: The loss curve during the training

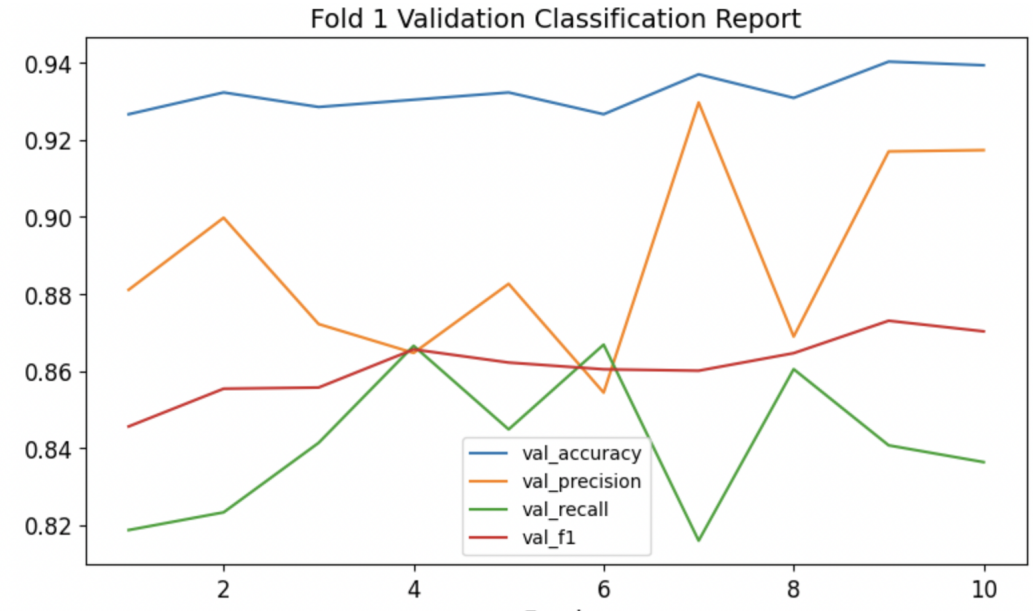

Figure 4.7: BERT using Stratified K-Fold 1: The curve of validation Accuracy, Precision, Recall and F-1 Score

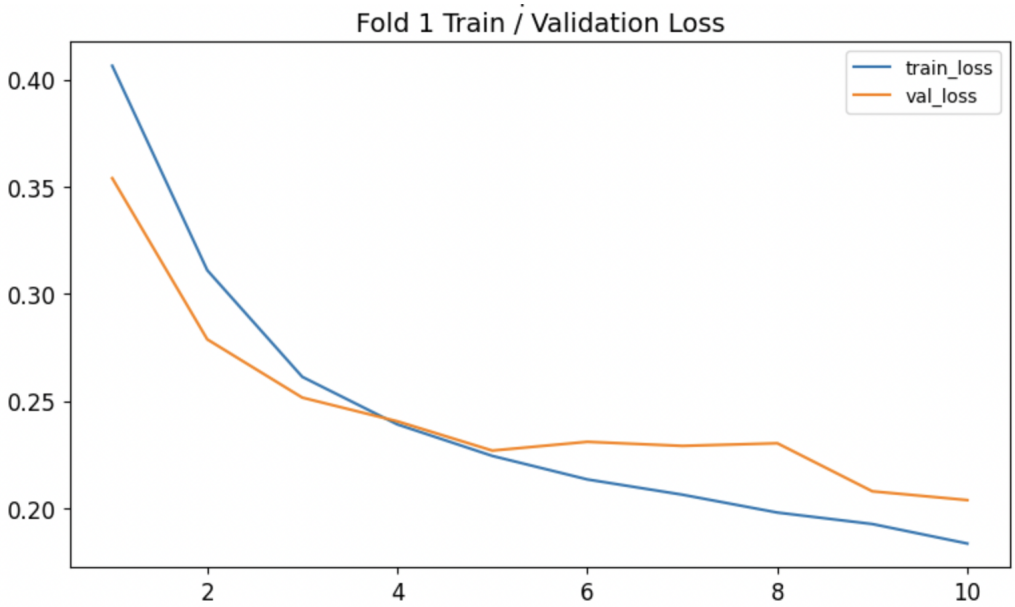

Figure 4.8: BERT using Stratified K-Fold 1: The loss curve during the training

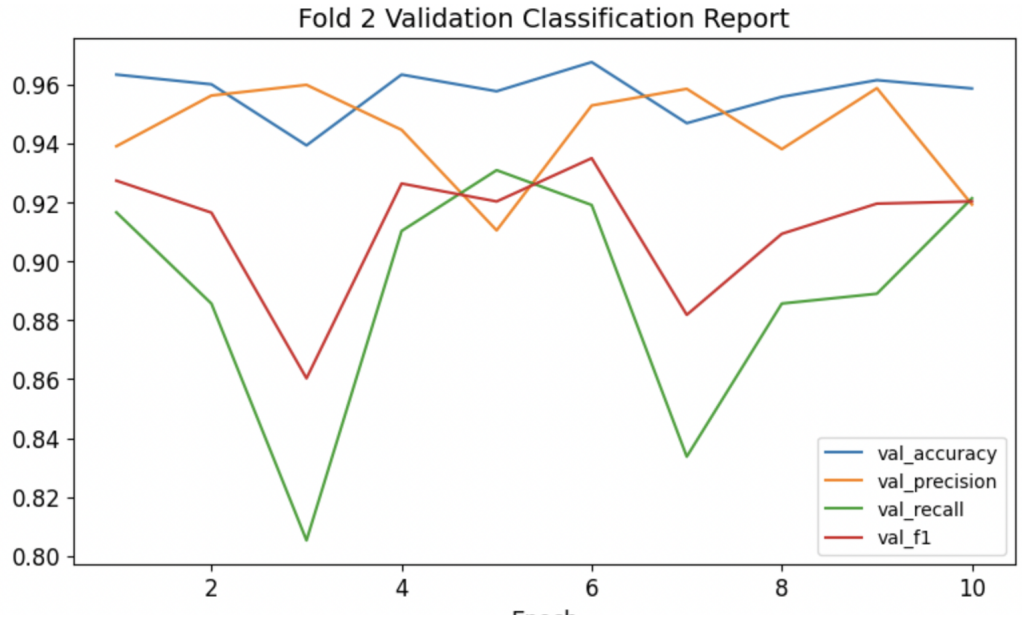

Figure 4.9: BERT using Stratified K-Fold 2: The curve of validation Accuracy, Precision, Recall and F-1 Score

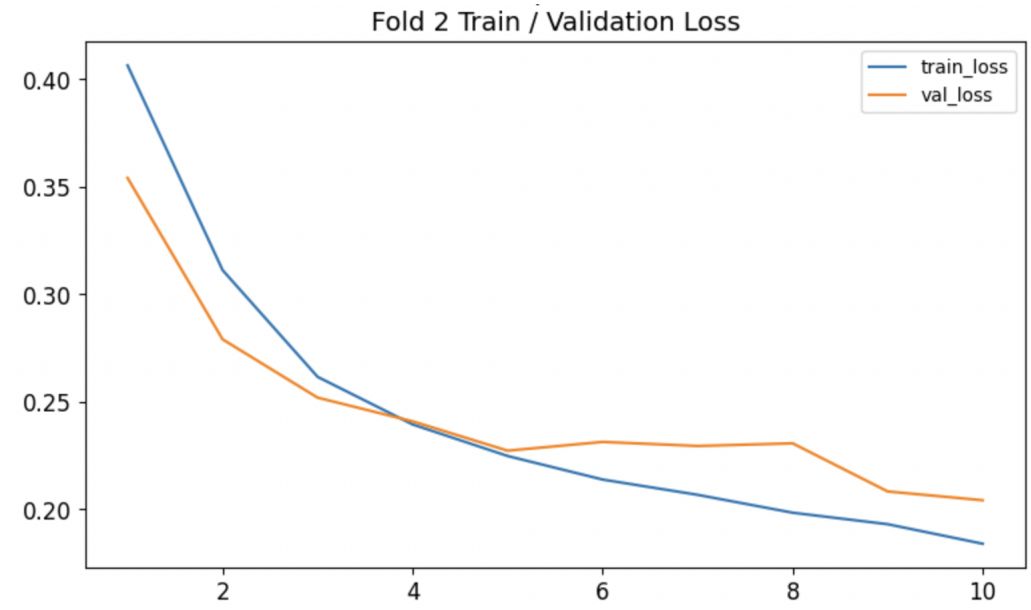

Figure 4.10: BERT using Stratified K-Fold 2: The loss curve during the training

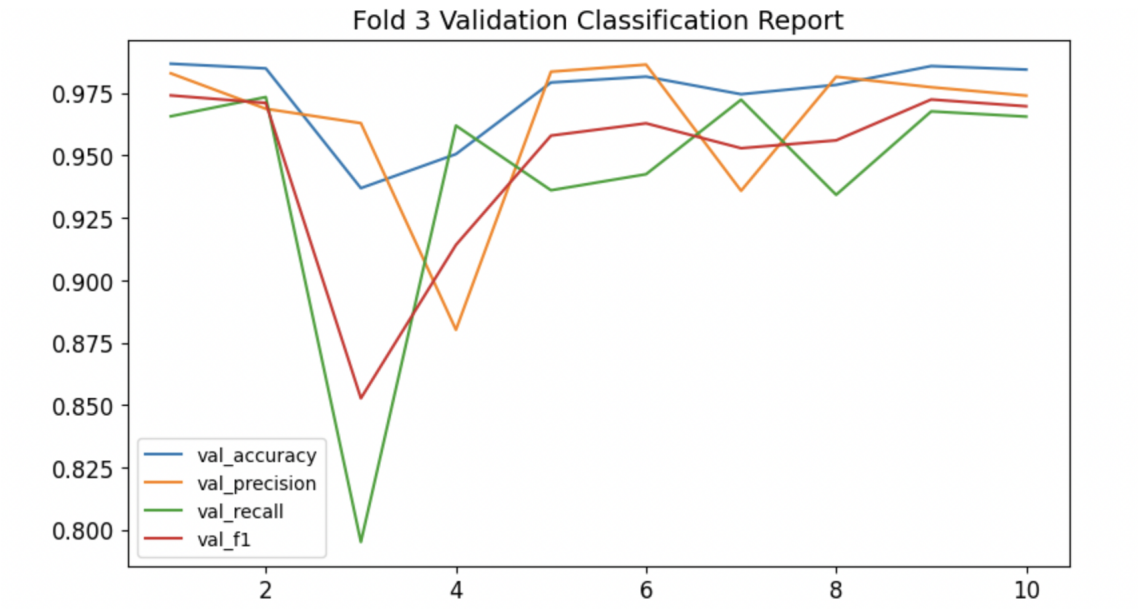

Figure 4.11: BERT using Stratified K-Fold 3: The curve of validation Accuracy, Precision, Recall and F-1 Score

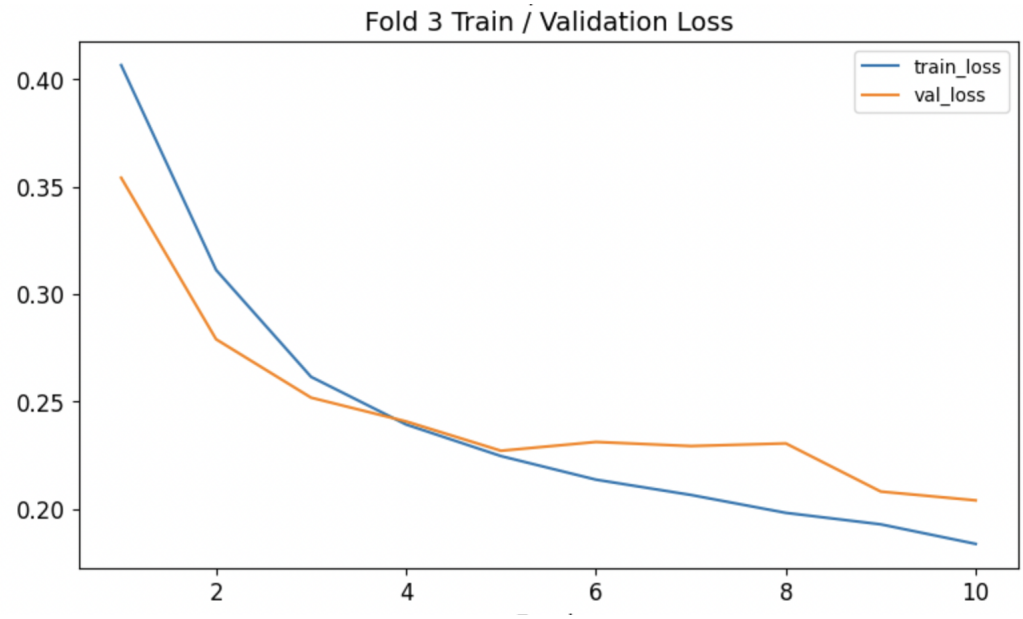

Figure 4.12: BERT using Stratified K-Fold 3: The loss curve during the training

4.4 Confusion Matrix Analysis

The confusion matrix is an important element in a machine learning model which determines the performance of the classification model. Accuracy, precision, responsiveness and recall values are all calculated using it. The divisions of true positive, false negative and true negative, false positive often take up the majority of the confusion matrix. Figure 4.13 demonstrates the evaluation of the confusion matrix, which compares the actual target values to the predictions provided by our model to show how well the classification model performs.

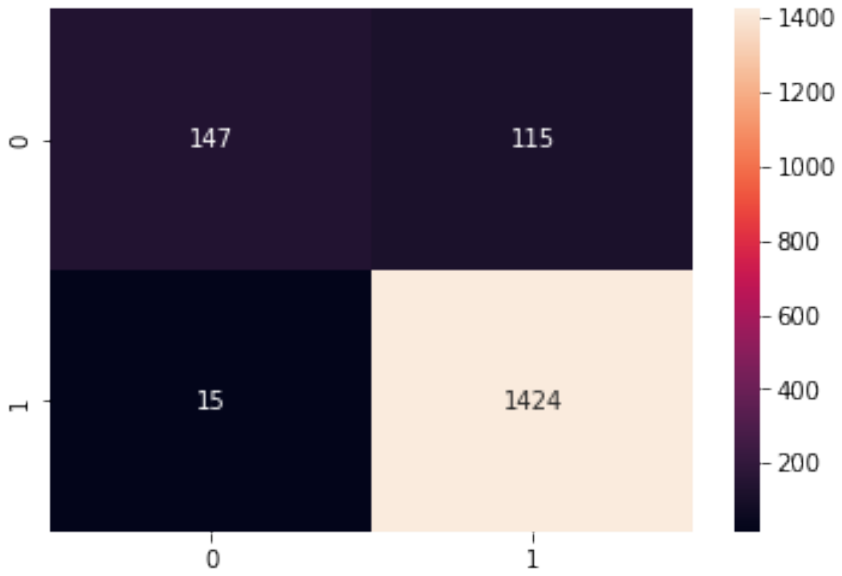

Figure 4.13: Confusion Matrix

4.5 Local Interpretable Model-Agnostic Explanations

LIME has been applied to add predictability to our data. Figure 4.14 and 4.15 shows some inputs and their LIME predictions whether they are considering as authentic or Fake news. Here, orange highlighted regions represent authentic news of a single word, whereas blue highlighted regions are represented as fake news.
Firstly, figure 4.14 depicts LIME confidently determines that input A is authentic news with 100% accuracy, and it emphasizes the key phrases so that users may see how LIME reached its conclusion. Secondly, input B is predicted as authentic news with an accuracy of 98%, and the basis for this prediction is also emphasized. While analyzing input C and D, our proposed model identified authentic news with 100% accuracy. Lastly, input E was classified as authentic news with 89% accuracy since LIME highlighted both it's positive and negative features, where the orange portion is regarded to be the authentic news and the blue portion is considered to be the fake news.
Secondly, Figure 4.15 shows another 5 inputs in addition to LIME predictions. LIME predicts fake news with 80% accuracy in input A and C and highlights the terms to illustrate its conclusion. Again, the prediction possibilities in input B demonstrate that the news is fake with an 84% accuracy, along with LIME predictions. Input D, on the other hand, predicts fake news with 99% accuracy, along with the highlighted key phrases. Finally, input E was classified correctly as fake news with 98% accuracy, and LIME has highlighted the significant phrases here as well.

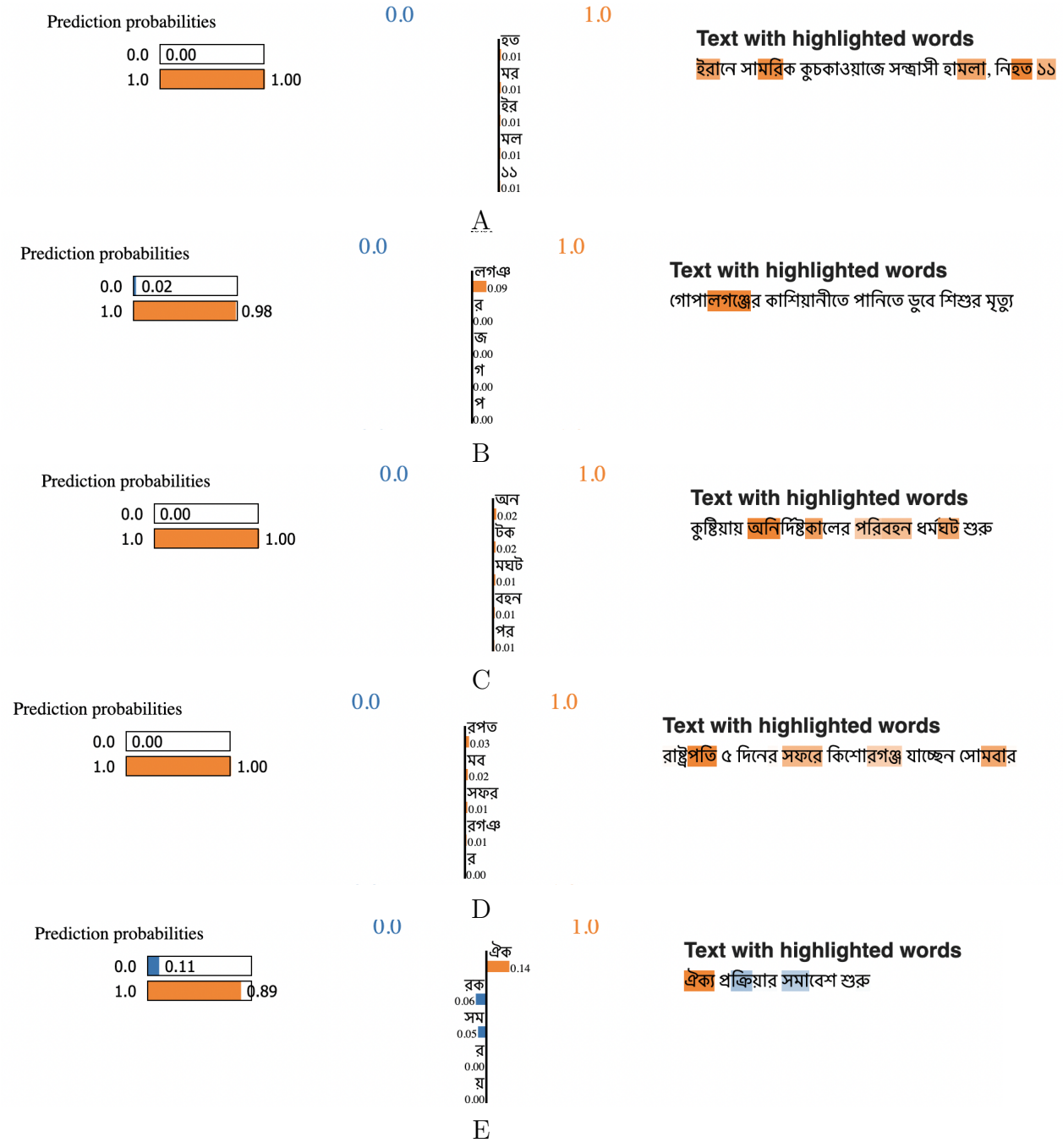

Figure 4.14: LIME Prediction: Authentic News. Here, Orange highlighted regions represent the Authentic News of a single word whereas Blue highlighted regions are represented as Fake News.

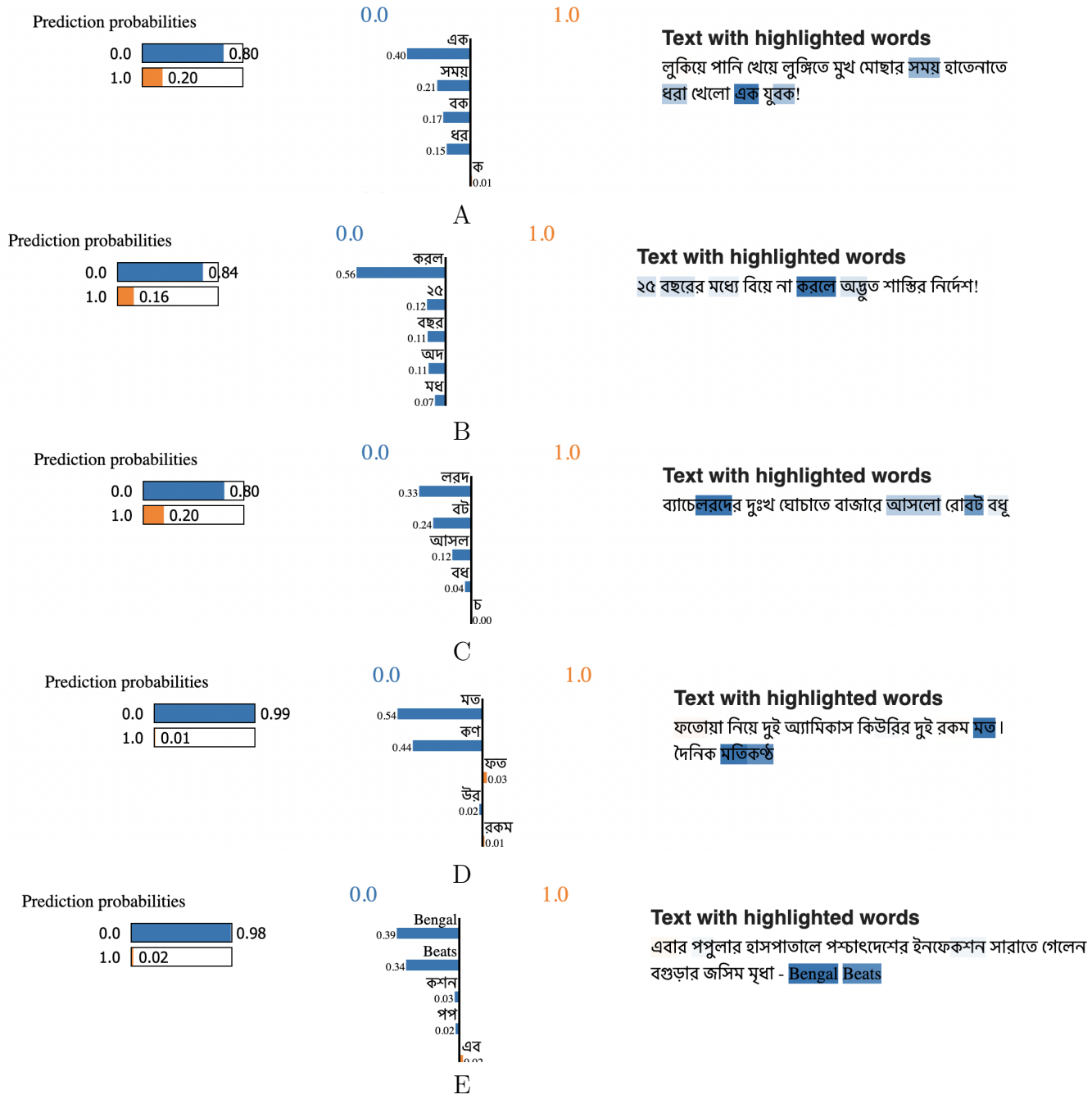

Figure 4.15: LIME Prediction: Fake News. Here, Orange highlighted regions represent the Authentic News of a single word whereas Blue highlighted regions are represented as Fake News.

4.6 Findings

Figure 4.16 demonstrates a comparison analysis of Random Forest and BERT with stratified K-fold (fold 3) considering different perimeters. In terms of precision, BERT has a precision of 97.40%, which is higher than the Random Forest model's precision of 86.48%. However, the recall value indicates an opposite situation, with the Random Forest model having a greater recall value of 100% than the BERT with stratified K-fold model, which has a recall value of 96.60%. Given that the accuracy scores for the two models are 98.45% and 86.83%, respectively, BERT with stratified K-fold appears to have done significantly better than Random Forest. Similarly, in terms of F1-Score, BERT with stratified K-fold outperforms Random Forest; hence, we consider BERT with stratified K-fold (fold 3) to be the top performing model in terms of detecting fake news in a given news headline dataset, with an F1-Score of 96.97%.

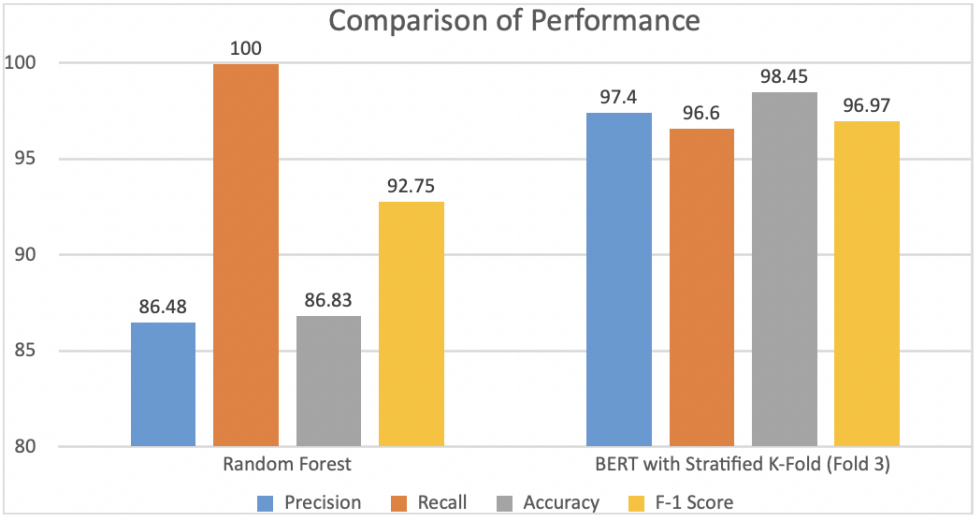

Figure 4.16: Comparison Analysis of Random Forest and BERT with Stratified K-Fold

Chapter 5

Conclusion

Fake news is a popular approach to attract reader's attention. Fake news spreads unfounded rumors and outright lies, filling people's minds with false information. It's similar to propaganda, and it's often used to persuade people to change their minds about particular topics. It makes use of exaggeration and, on occasion, outright falsification. Surprisingly, fake news spreads quicker than any infection. Fake News can influence a nation's social or economic equilibrium on a personal and global level. There are numerous drawbacks to fake news, which can spread to a wider level, if not addressed quickly using a system like ours, which can be utilized to mitigate the issues listed above to a greater extent. Therefore, building a model to detect fake news is crucial to protect the integrity of the media in a nation and save the society from the chaos of misleading information. We provided a comparison of Machine Learning models in this study to select the best fitting model, with Random Forest which attaining the maximum accuracy of 86.48 percent. Furthermore, we employed a BERT-based model with stratified K-Fold cross validation to detect Bangla fake news with 98.45 percent accuracy with the validation data. Additionally, we have added LIME, which adds interpretability to our system. BERT's pre-trained and fully linked layers are used in this research to collect deep properties and classify false news. In future work, we want to collect more relevant datasets and test out different algorithms. We also would like to use SHAP (Shapley Additive Explanations) explainable AI to interpret fake news and rumors.

Bibliography

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

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

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

[4] T. Sraboni, M. Uddin, F. Shahriar, R. A. Rizon, S. I. S. Polock, *et al.*, "Fakedetect: Bangla fake news detection model based on different machine learning classifiers," Ph.D. dissertation, Brac University, 2021.

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

[6] J. M. Burkhardt, "History of fake news," *Library Technology Reports*, vol. 53, no. 8, pp. 5–9, 2017.

[7] G. Di Domenico, J. Sit, A. Ishizaka, and D. Nunan, "Fake news, social media and marketing: A systematic review," *Journal of Business Research*, vol. 124, pp. 329–341, 2021.

[8] S. I. Manzoor, J. Singla, *et al.*, "Fake news detection using machine learning approaches: A systematic review," in *2019 3rd international conference on trends in electronics and informatics (ICOEI)*, IEEE, 2019, pp. 230–234.

[9] 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.

[10] C. Elliott, *Here are the real fake news sites*, https://www.forbes.com/sites/christopherelliott/2019/02/21/these-are-the-real-fake-news-sites/?sh=63baf7913c3e, 2019.

[11] S. Hölig, U. Hasebrink, and J. Behre, *Reuters institute digital news report 2019: Ergebnisse für Deutschland*. DEU, 2019, vol. 47.

[12] F. Karim, "Fake news on social media–who consume it and why: Bangladesh perspective," *Communication and Media in Asia Pacific (CMAP)*, vol. 4, no. 1, pp. 11–22, 2021.

[13] *Busting the top 3 fake news of the week*, https://www.tbsnews.net/thoughts/busting-top-3-fake-news-week-173236, 2020.

[14] "Mobs beat 2 for 'kidnapping'," *The Daily Star*, 2019. [Online]. Available: https://www.thedailystar.net/frontpage/news/mobs-beat-2-dead-kidnapping-1774471.

[15] M. G. Hussain, M. R. Hasan, M. Rahman, J. Protim, and S. A. Hasan, "Detection of bangla fake news using mnb and svm classifier," *arXiv preprint arXiv:2005.14627*, 2020.

[16] R. R. Mandical, N. Mamatha, N. Shivakumar, R. Monica, and A. Krishna, "Identification of fake news using machine learning," in *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, IEEE, 2020, pp. 1–6.

[17] H. Bingol and B. Alatas, "Rumor detection in social media using machine learning methods," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, IEEE, 2019, pp. 1–4.

[18] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity*, vol. 2020, 2020.

[19] A. Al Imran, Z. Wahid, and T. Ahmed, "Bnnet: A deep neural network for the identification of satire and fake bangla news," in *Computational Data and Social Networks*, S. Chellappan, K.-K. R. Choo, and N. Phan, Eds., Cham: Springer International Publishing, 2020, pp. 464–475, ISBN: 978-3-030-66046-8.

[20] F. Harrag and M. K. Djahli, "Arabic fake news detection: A fact checking based deep learning approach," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 21, no. 4, Jan. 2022, ISSN: 2375-4699. DOI: 10.1145/3501401. [Online]. Available: https://doi.org/10.1145/3501401.

[21] E. Masciari, V. Moscato, A. Picariello, and G. Sperli, "A deep learning approach to fake news detection," in *International Symposium on Methodologies for Intelligent Systems*, Springer, 2020, pp. 113–122.

[22] *Study finds fake news spreads faster than real news on twitter*, shorturl.at/BJKN0, 2018.

[23] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid cnn-rnn based deep learning approach," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100 007, 2021.

[24] M. Z. Hossain, M. A. Rahman, M. S. Islam, and S. Kar, "Banfakenews: A dataset for detecting fake news in bangla," *arXiv preprint arXiv:2004.08789*, 2020.

[25] E. Hossain, N. Kaysar, J. U. Joy, A. Z. Md, M. Rahman, W. Rahman, *et al.*, "A study towards bangla fake news detection using machine learning and deep learning," in *Sentimental Analysis and Deep Learning*, Springer, 2022, pp. 79–95.

[26] F. Islam, M. M. Alam, S. S. Hossain, *et al.*, "Bengali fake news detection," in *2020 IEEE 10th International Conference on Intelligent Systems (IS)*, IEEE, 2020, pp. 281–287.

[27] T. Islam, S. Latif, and N. Ahmed, "Using social networks to detect malicious bangla text content," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 2019, pp. 1–4. DOI: 10.1109/ICASERT.2019.8934841.

[28] M. Z. H. George, N. Hossain, M. R. Bhuiyan, A. K. M. Masum, and S. Abujar, "Bangla fake news detection based on multichannel combined cnn-lstm," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, 2021, pp. 1–5.

[29] P. B. Pranto, S. Z.-U.-H. Navid, P. Dey, G. Uddin, and A. Iqbal, "Are you misinformed? a study of covid-related fake news in bengali on facebook," *arXiv preprint arXiv:2203.11669*, 2022.

[30] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[31] J. R. Quinlan, "Learning decision tree classifiers," *ACM Computing Surveys (CSUR)*, vol. 28, no. 1, pp. 71–72, 1996.

[32] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, "An introduction to decision tree modeling," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 18, no. 6, pp. 275–285, 2004.

[33] Y.-Y. Song and L. Ying, "Decision tree methods: Applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.

[34] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009.

[35] L. Jiang, S. Wang, C. Li, and L. Zhang, "Structure extended multinomial naive bayes," *Information Sciences*, vol. 329, pp. 346–356, 2016.

[36] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial naive bayes for text categorization revisited," in *Australasian Joint Conference on Artificial Intelligence*, Springer, 2004, pp. 488–499.

[37] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed, "Multinomial naive bayes classification model for sentiment analysis," *IJCSNS Int. J. Comput. Sci. Netw. Secur*, vol. 19, no. 3, p. 62, 2019.

[38] C. Mood, "Logistic regression: Why we cannot do what we think we can do, and what we can do about it," *European sociological review*, vol. 26, no. 1, pp. 67–82, 2010.

[39] J. M. Hilbe, "Logistic regression.," *International encyclopedia of statistical science*, vol. 1, pp. 15–32, 2011.

[40] L. Bottou, "Stochastic gradient descent tricks," in *Neural networks: Tricks of the trade*, Springer, 2012, pp. 421–436.

[41] N. Ketkar, "Stochastic gradient descent," in *Deep learning with Python*, Springer, 2017, pp. 113–132.

[42] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[43] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine learning*, Elsevier, 2020, pp. 101–121.

[44] 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.

[45] S. Mishra, B. L. Sturm, and S. Dixon, "Local interpretable model-agnostic explanations for music content analysis.," in *ISMIR*, vol. 53, 2017, pp. 537–543.

[46] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[47] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *European conference on information retrieval*, Springer, 2005, pp. 345–359.

[48] D. Fourure, M. U. Javaid, N. Posocco, and S. Tihon, "Anomaly detection: How to artificially increase your f1-score with a biased evaluation protocol," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2021, pp. 3–18.

[49] Z. H. Hoo, J. Candlish, and D. Teare, *What is an roc curve?* 2017.

[50] P. Flach and M. Kull, "Precision-recall-gain curves: Pr analysis done right," *Advances in neural information processing systems*, vol. 28, 2015.

[51] J. T. Townsend, "Theoretical analysis of an alphabetic confusion matrix," *Perception & Psychophysics*, vol. 9, no. 1, pp. 40–50, 1971.

[52] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06, Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, pp. 233–240, ISBN: 1595933832. DOI: 10.1145/1143844.1143874. [Online]. Available: https://doi.org/10.1145/1143844.1143874.