Bachelor of Science in Computer Science & Engineering

BDFN: A Bilingual Model to Detect Online Fake News Using Machine Learning Technique

by

Fahmida Liza Piya

ID: 1504107

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

April, 2021

BDFN: A Bilingual Model to Detect Online Fake News Using Machine Learning Technique

Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Fahmida Liza Piya

ID: 1504107

Supervised by

Dr. Mohammad Shamsul Arefin

Professor

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

The thesis titled '**BDFN: A Bilingual Model to Detect Online Fake News Using Machine Learning Technique'** submitted by ID: 1504107, Session 2019-2020 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

Supervisor

Dr. Mohammad Shamsul Arefin
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (Ex-Officio)

Dr. Asaduzzaman
Professor & Head
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (External)

Dr. Mohammad Moshiul Hoque
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and that neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Signature of the candidate

Date: 19.04.2021

Acknowledgements

The success and final result of this thesis involved a great deal of support and assistance from many people, and I consider myself incredibly fortunate to have received it all the way through my thesis. Professionally and socially, it has been a rewarding experience. Much of what I've accomplished has been possible only because of such oversight and assistance. I've dabbled in a variety of fields related to my profession and have learned a lot.

All praise and thanks to Almighty Allah, the most merciful and generous, for allowing me to complete this project and report. I would also like to show my greatest gratitude and honour to Prof. Dr. Mohammad Shamsul Arefin sir for his continuous direction, support and encouragement. His constant push towards success helped me out extremely throughout the process. His continuous guidance and encouragement empowered me to continue working on my thesis even during the most critical period of the pandemic and to successfully complete it.

Our department, the Department of Computer Science and Engineering, has always been a source of inspiration, support, and guidance for me. Thank you to our department head Prof. Dr. Asaduzzaman for supporting us with all of the necessary resources.

Finally, I'd like to express my gratitude to my parents for their unwavering compassion and understanding as I traversed the ups and downs of my undergraduate studies.

Abstract

The dissemination of online misinformation is causing increasing concern around the world. Government officials and other responsible agencies are deeply worried about the situation and are working tirelessly to change it, but it seems that some yellow journalists are only interested in making money by selling news with clickbait headlines and fake facts inside the news. Fake news spreads faster than true news because people are more likely to post fake news than real news. Some shady online news sources, both in English and Bengali, are actively working to spread false information in order to create a stir. We propose a bilingual fake news detection model in this paper that employs TF-IDF and N-Gram Analysis for feature extraction in order to detect fake news from a bilingual perspective. In addition, we compare the results of six separate machine learning algorithms for detecting false news. The model employs a supervised method of operation. Among these, we have acquired the highest performance with Linear Support Vector Classification (Linear SVC) algorithm where the accuracy is 93.29% and the F1 score is 0.93.

Keywords— Fake News, Natural Language Processing, Bilingual, Text Analysis, TF-IDF, Text Analysis, N-Gram Models, Online News Portals.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

clickbait Information on the internet with the primary goal of attracting interest and encouraging users to click on a link to a specific web page. iv

kernel The term "kernel" is used because the Support Vector Machine uses a collection of mathematical functions to provide a window to manipulate data. As a result, the Kernel Function transforms the training set of data in such a way that a non-linear decision surface can be converted into a linear equation in a larger number of dimension spaces. 23

prior probability This probability can be characterized as prior knowledge or belief, or the probability of an event calculated before new data is collected. As new information becomes available, this probability is updated to provide more reliable results. 26

Chapter 1

Introduction

1.1 Introduction

Fake news is described as false stories that appear to be news and are widely disseminated on the internet or by other media, typically to sway political or other opinions or as a joke. Because of its potential to have harmful consequences, the spread of fake news and its dissemination on various platforms has become a major concern.

Prior to the introduction of the Internet, journalists were charged with checking reports and sources and thoroughly checking the details [1]. Therefore, the public's access to fake news were more limited. Today, social media enables the distribution of unverified and inaccurate knowledge to a larger extent, thereby deeply impacting the perception of the readers and interpretation of the events [2]. As a normal human action, authors [3] stressed the lack of interest in checking the authenticity of much of the knowledge found in blogs and virtual encyclopedias, and the inherent propensity to trust in the shared content.

Fake news is on the rise these days, particularly news about Covid-19, but most people can't tell the difference between real and fake news. The proliferation of "fake news" has intensified political polarization, reduced confidence in public institutions, and weakened democracy [4]. In the last decade, there has been a dramatic rise in the dissemination of fake news, most notably during the 2016 US elections [5]. Such a flood of posting stories online that do not adhere to reality has resulted in a slew of issues, not only in politics, but in sports, health, and science as well [6]. The financial markets [7] are one sector affected by false news, where a rumor can have negative effects and even bring the economy to a halt.

Some news portals deliberately publish fake news and by this they harm the reputation of one, might even spoil someones career. Journalist create clickbait to add revenue to their pocket. Recently, the GOVT. also faced major problems because of spreading rumors and fake news. Without any kind of guidelines, only spending a very small amount of money, anyone can be an editor of a website and most of the time the websites are maintained by one single person only. They are not aware of the most basic journalism ethics, dont abide by the rules and do not even check the basic facts before posting a news. Some of the portals only concentrate on getting viral to get benefited financially.

Detecting fake news presents a number of fresh and difficult study issues. In this paper, we present a model that can detect fake news from a bilingual corpus which means a corpus which contains both Bengali and English language together. And to accomplish this novel task we have chosen the text analysis techniques so that we can extract and classify information from the contents of a news. Here, we have analyzed both the headline and the main text body of a specific news in both Bengali and English. Logistic Regression, Linear SVC, Decision Tree Classifier, Random Forest Classifier, Multinomial Naive Bayes and Passive Aggressive Classifier were compared as classification techniques to recommend the best model for our fake news detection system.

1.2 Framework/Design Overview

Concern about the issue is widespread. However, there is still a lot of mystery surrounding the vulnerabilities of individuals, organisations, and culture to malicious actors. A new safeguarding framework is needed. To detect false news, various machine learning methods have been proposed. Preprocessing the dataset is the first step in building our bilingual model. Initially, we have two separate datasets. The first move is to do the pre-processing. We tokenized them with NLTK, stemmed them with Porter Stemmer, and then eliminated stopwords.

We used the N Gram model after cleaning the text corpus. This model has been incorporated into our classification technique because it significantly improves

precision. This method extracts bigrams, trigrams, and even four grams to compare and test the accuracy of the fake news detection model. TF-IDF is now applied to these word sequences.

Our pre-processed data collection was labelled. The dataset is split into a 70:30 ratio, implying that we used 70% of our labeled dataset to train our classifier, allowing it to later classify any news as real or false based on its previous training experience with N-Gram features and TF-IDF ratings.

When detecting any news, the classifier will take into account the highest feature values. For our bilingual fake news detection model, we used six separate machine learning algorithms. They are: i) Logistic Regression ii) Linear Support Vector Classification iii) Decision Tree iv) Random Forest Classifier v) Multinomial Naive Bayes vi) Passive Aggressive.

1.3 Difficulties

The proposed work presented a variety of new challenges. While fake news can be found in both Bengali and English news stories, and we needed a bilingual dataset to identify and train on it, finding one was difficult.

- Although there are numerous benchmark datasets for detecting fake news in English from various perspectives. A Bengali-annotated dataset was only recently published.
- However, because our model needs bilingual vision training, we had to merge two datasets as our final dataset, which eventually took a long time and effort.
- We also started creating our own dataset initially, but we had to stop because there were no automated annotation tools for distinguishing fake and authentic news. We also couldn't form a data annotation team to manually review the labels, and that is a difficult task for a single person to complete.
- The pre-processing part was a bit time-consuming as we had to clean data

in a bilingual perspective. Separate tools were needed to clean Bangla and English data and to finally combine them as our final dataset.

1.4 Applications

People are inundated with news on a daily basis, but they lack the ability to distinguish between true and false news. These are often used to deceive and/or trick the general public into believing something that is not true. Even if mis -information or false news is corrected, the negative effect on an individual or society as a whole can last a lifetime. Our architecture can be used to create systems that automatically detect fake news and filter it out regardless of whether the language is Bengali or English, preventing it from spreading.

1.5 Motivation

Several content based solutions have been proposed in recent years based on text mining approach but there is a problem that denotes most of them were confined to one specific language only. Till now, most of the work done on fake news detection revolves around the language English like the work in [8], also there is a work done in both Chinese and English language [9], another one in Spanish [10] and in Bengali [11]. However, fake news detection from online news is still in the early stages of growth, and many difficult problems still need to be investigated further. And previously these works have been related to fake news detection specifically in English or Bengali but very few were done in bilingual approach. According to our research, not a single work has been done that has the ability to detect fake news from a context that can combine both Bengali and English language. A novel model like this can help in advanced implementations where fake news can be identified regardless of whether the text in the news is in Bengali or English.

1.6 Contribution of the thesis

The aim of a thesis or research project is to achieve a particular set of objectives, such as defining a new approach or improving an existing one. A thesis or research project should add to the advancement of human knowledge by providing new information. In this thesis of ours, the main focus was to develop e framework for detecting fake news in a bilingual perspective.

- Feature extraction from bilingual dataset using TF-IDF along with N-Grams.
- To identify fake news and authentic news from different news articles of online news portals from a bilingual perspective.
- Apply text analysis technique and compare six different machine learning techniques to compare performance.

1.7 Thesis Organization

The following is how the rest of this study report is organized:

- Chapter 2 gives a short overview of existing related work on Fake News Detection.
- Chapter 3 discusses the proposed methodology to detect fake news from bilingual perspective. We have used N-Gram models with TF-IDF for performing feature extractions. And for performing the novel task we have implemented six machine learning models which are: Logistic Regression, Linear SVC, Decision Tree, Random Forest, Multinomial Naive Bayes and Passive Aggressive Classifier.
- Chapter 4 discusses the dataset which we have used for developing our model along with an overview of the performance measure for our proposed system.
- Chapter 5 contains the overall summary of this the thesis work and provides some future directions and recommendations as well.

1.8 Conclusion

The widespread dissemination of fake news has the potential to damage both individuals and the society. For starters, fake news has the potential to disrupt the news ecosystem's credibility balance. Second, fake news is designed to convince users to believe in misleading or false information. Propagandists also use fake news to spread political agendas or exert influence. Third, fake news alters how people perceive and respond to real news.

It's important that we build a system to automatically identify fake news on social media to help mitigate the harmful impact of fake news, both for the good of the public and the news ecosystem. Our architecture which is built using machine learning models can be used to build systems that will detect fake news automatically, regardless of whether it is written in Bengali or English, preventing it from spreading on internet and make it a better place.

Chapter 2

Literature Review

2.1 Introduction

Since this is a relatively new phenomenon, at least in terms of societal interest, research on bilingual fake news identification is still in its early stages. There have been quite a lot of work using different machine learning techniques. The following is a summary of some of the published work related to our work on fake news detection.

2.2 Related Literature Review

A straightforward strategy was suggested to detect fake news in social media networks using certain data from articles such as title, use of title numbers, capital letters in title, on-site user behaviour, occurrence of title phrases in content and title keywords [12].

Authors [13] suggested that machine learning techniques can be used to detect fake news and three common approaches are used in their experiments: Naive Bayes, Neural Networks, and Support Vector Machines. The normalization process, according to this study, is an essential step in cleaning data before using the machine learning method to classify it.

A survey was performed by authors [14] analyzing different fake news assessment methods based on linguistic methods and network analysis approaches. In another work authors [15] used naive Bayes classifier which is a simple approach for detecting fake news which was tested on some news posts from Facebook.

In an article [16], the fake news problem is explored by authors by conducting

a two-phase analysis of current literature: characterization and detection. They presented the basic concepts and principles of fake news in both conventional media platforms and social media during the characterization process. They examined current fake news detection approaches from a data mining perspective, including feature extraction and model creation, during the detection process. They have also discussed the datasets, evaluation metrices and promising future directions in this specific field.

Authors [17] have identified clickbaits and installed browser extensions to warn and prevent users from accessing these type of news with clickbait headlines. To identify false or true claims in social media, websites and other news sources authors have taken the interaction between the language of the claim and the web sources reliability jointly in a work [18]. In [19], dishonest opinions were identified by using applied stylometric tools, i.e. lexical and syntactic; Sequential Minimal Optimization (SMO), Naive Bayes and Support Vector Machine (SVM).

N-Gram analysis for a fake news detection model have been used and machine learning techniques have also been used [20]. In this paper, they have got the highest accuracy of 92% using Linear Support Vector Machine (LSVM) which is a classifier and along with this TF-IDF has been used as the feature extractor. Wang [21] took a slightly different approach. For training various ML models, the author used textual features and metadata. The author concentrated on convolutional neural networks (CNN). To capture the dependency between the metadata vectors, a convolutional layer is used, followed by a bidirectional LSTM layer.

In another machine learning-based framework for detecting fake news has been developed, which uses term frequency-inverse document frequency (TF-IDF) of bag of words and n-grams as a feature extraction technique and Support Vector Machine (SVM) as a classifier [22].

Taking into account the content of online news stories, the authors in their work [23] suggested a fake news identification method and investigate two machine learning algorithms using word and character n-grams analysis. Their tests with character n-grams and the Term-Frequency-Inverted Document Frequency

(TF-IDF) and Gradient Boosting Classifier produce better performance, with a 96 percent accuracy [23].

Another machine learning-based fake news identification model employs an ensemble approach [24]. Multiple learning algorithms are used in ensemble methods in statistics and machine learning to achieve greater predictive precision than any of the constituent learning algorithms alone. The research looks into various textual properties that can be used to tell the difference between fake and real content. Authors trained a combination of different machine learning algorithms using various ensemble methods and tested their output on four real-world datasets using those properties. In contrast to individual learners, the proposed ensemble learner method performed better in experiments.

In a recent work [25] on bangla fake news detection system authors have analyzed their recently released Bangla dataset and created a benchmark framework using cutting-edge NLP techniques to detect Bangla fake news. They [25] have used conventional linguistic features as well as neural network-based approaches to develop this framework.

2.3 Conclusion

The above mentioned works mostly focus on detecting fake news from one single language using modern cutting edge technologies and found great success most of the time. Although the base of our framework was generated by following the already existing literature covering different machine learning techniques. However, developing a bilingual system is still in its infancy, and only a few works address this subject.

2.3.1 Implementation Challenges

There are a few challenges which we faced during the implementation of our models which are listed below:

- Despite the fact that online news can be gathered from a variety of sources,

manually determining the veracity of news is a difficult task that typically requires domain experts to conduct a thorough review of statements and additional facts, background, and reports from authoritative sources. Due to these difficulties, existing public databases of fake news specially in English language are very small.

- The pre-processing of the Bilingual dataset was a bit difficult as we had to learn new tools like NLTK.
- It was very time consuming after we decided to work with N-Grams and explored our dataset even using the Four-Grams.
- After the feature extractors we decided to compare six different machine learning classifiers and after implementing all these using python the entire model needed a long time to perform compilation and generate the performance results.

Chapter 3

Methodology

3.1 Introduction

To develop our bilingual fake news detection model we have developed a framework first. After deciding on our dataset which actually covered the majority of the prerequisite step the pre-processing of text dataset was performed. After cleaning the dataset features extractors are used to extract features from our bilingual text dataset. Combination of features and labels are fed into our training models to get familiarized with our labeled dataset. Then after completion of training the model is ready to perform test operations on some unlabeled data and identify fake and authentic news based on it's previous experience.

3.2 Diagram/Overview of Framework

The first step towards creating our bilingual model is to preprocess the dataset. We have two different datasets initially. The pre-processing step is conducted at first. We have tokenized them using NLTK, conducted stemming using porter stemmer then removed stopwords. We uploaded a list of Bengali stopwords in our python environment while removing stopwords from the Bengali newspaper dataset. And, the English stop words have been removed using NLTK. Then punctuations and garbage characters have been removed from the dataset.

After cleaning the text corpus we implemented the N Gram model. We have integrated this model in our classification technique as it accelerates the accuracy greatly. Bigrams, Trigrams and even four grams are extracted in this process to compare and evaluate the accuracy of the fake news detection model. Now taking these sequences of words TF-IDF is applied. The last step is the training

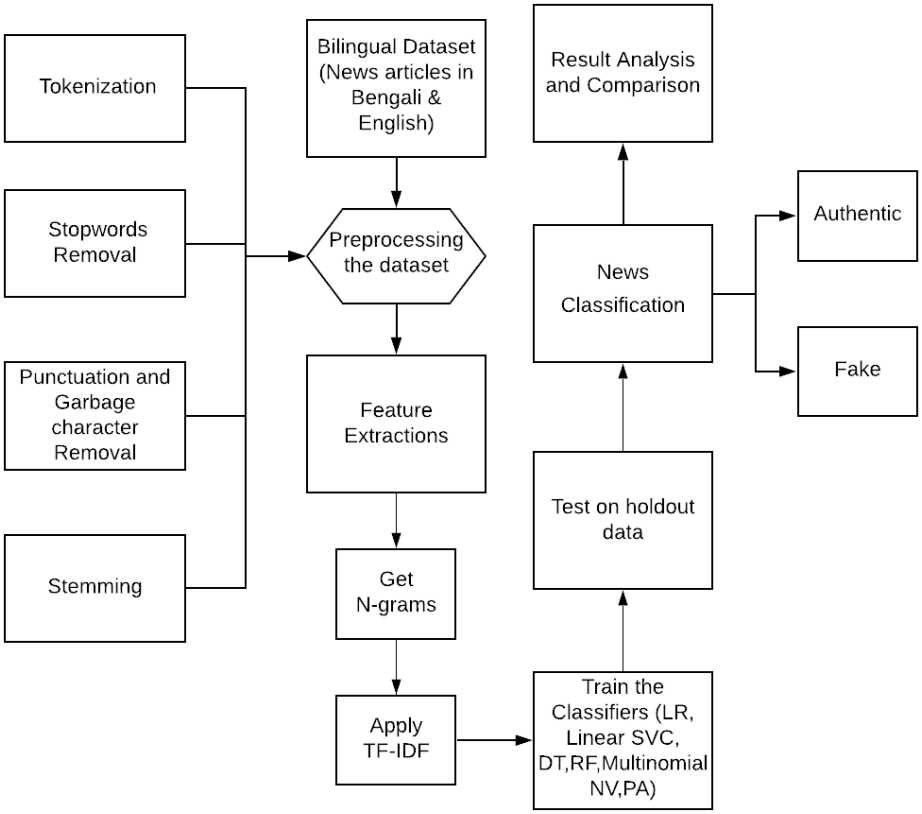

Figure 3.1: Preprocessing and classification process of our bilingual model for fake news detection.

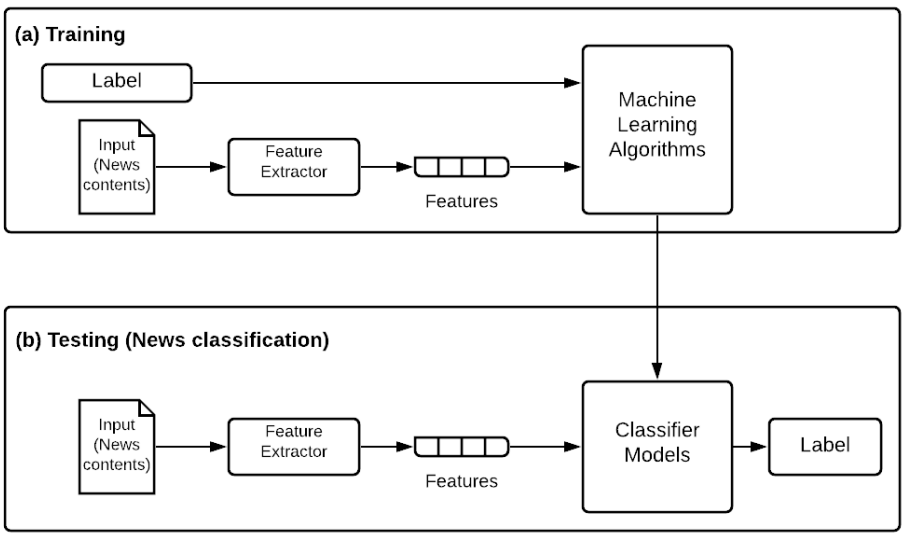

Figure 3.2: Fake News classification Technique.

Title	Text	Category	Label
হঠাৎ আফগান ক্রিকেট বোর্ড প্রধানের পদত্যাগ	ক্রিকেট বিশ্বের নতুন চমকের নাম আফগানিস্তান। কয়েক বছরে তাদের পারফরম্যান্স ...	Sports	1
বিশ্বের ১৭৩ জনের মধ্যে ৩য় সৎ সরকার প্রধান শেখ হাসিনা	পিপলস অ্যান্ড পলিটিক্স বিশ্বের পাঁচজন সরকার রাষ্ট্রপ্রধানকে চিহ্নিত করেছেন, যাদের দুর্নীতি স্পর্শ করেনি, বিদেশে কোনো ব্যাংক অ্যাকাউন্ট নেই....	National	0
Muslims in Asia protest against Trump's Jerusalem plan	Thousands of protesters in Muslim-majority countries in Asia rallied on Friday to condemn the U.S. decision to recognize	worldnews	1
Donald Trump Sends Out an Embarrassing Message!	Donald Trump Sends Out Embarrassing New Years Eve Message; This is Disturbing	News	0

Figure 3.3: Bilingual Dataset after pre-processing

of the classifier. Our dataset that has been pre-processed is labeled. The dataset is splitted into 70:30 ratio which means we have used 70% of our labeled dataset for training our classifier so that later on it can identify any news as true or fake based on its previous training experience using N-Gram features and TF-IDF scores.

In fig. 3.2 (a) During training our bilingual model, feature extractor (Tf-Idf and N-Gram model) is used to convert each input news contents into a feature matrix. Combinations of feature sets and the labels of the news contents are fed into machine learning algorithms. (b) While performing classification, the exact same feature extractors are used to identify feature sets. Features are fed into the machine learning models to perform classification of the news articles.

3.3 Detailed Explanation

The classifier will consider the highest feature values while detecting any news. We have used six different algorithms of machine learning for our bilingual fake news detection model. And, they are- i) Logistic Regression ii) Linear Support Vector Classification iii) Decision Tree iv) Random Forest Classifier v) Multinomial Naive Bayes vi) Passive Aggressive. For analyzing the text we have used both the news headline and the content of a specific news but as the contents had more words for analysis the models showed better performance.

3.3.1 Dataset Preparing

For our novel model, we needed a dataset that included both English and Bengali texts. So, after some significant processing, we merged two separate datasets. For the bengali texts we used a recently available public dataset [25]. Here, the authors have proposed a dataset for fake news detection containing 50K annotated data. There are 12 Categories of data. We have used 7k authentic news and 1k fake news from this specific dataset to build our model. The data that has been selected is labeled as we are approaching towards building a supervised model. The information columns that the dataset contains are- Label, domain Name, Published Time, Category, Source, Headline-article relation, Headline, Article. We have to drop some of the columns to match with our English Dataset.

In addition, for our English news content, we used a dataset that the authors published in [20]. This specific dataset is known as ISOT Fake news dataset. There are two types of articles-True and Fake. There are more than 12600 articles collected from Reuters which are true and also more than 12600 fake articles collected from various sources. Each article contains the following information: article title, text, type and the date. And to match with the Bengali dataset we labeled both the true and fake articles. The true articles are labeled as 1 and the fake articles are labeled as 0. To combine these two datasets in different languages we had to drop some of the columns and rename them. Then, we worked with Title, Text, Category, Date and Label.

- **Title**: Short title text that seeks to grab readers' attention and defines the article's main subject.
- **Text**: This is the main text which elucidates on the news story's information.
- **Category**: It shows the category of the news article. Ex: Sports, Politics etc.
- **Date**: Denotes when the story was published
- **Label**: Denotes if the news story is true or fake.

After combining our dataframe the data cleaning and preprocessing step has been performed. We have done text analysis of both the Title and the Text column.

3.3.2 Data Preprocessing

The process of converting data to something that a computer could understand is called pre-processing. Before applying any feature extractor or classifier model data processing is needed to be done so that we can get rid of unnecessary information existing in the data. And to apply the N-Gram model we want our dataset to be prepared in a specific way. Our first step is to apply tokenization to get tokens from the text of every news article.

3.3.2.1 Tokenization

To make our computer understand the text, we must break down the word in a way so that our machine can understand. Here, comes the concept of tokenization from Natural Language Processing (NLP). While working with text data tokenization is one of the key or mandatory aspects. If we dont perform tokenization we will not be able to work with text data.

In our bilingual dataframe, we have simply conducted splitting for further cleaning and preprocessing tasks. We have tokenized the texts of both the columns naming Headline and Text. Example of tokenized text is shown in fig. 3.4 & fig. 3.5.

Text	Tokenized Text
ক্রিকেট বিশ্বের নতুন চমকের নাম আফগানিস্তান	ক্রিকেট, বিশ্বের, নতুন, চমকের, নাম,আফগানিস্তান
এশিয়া কাপের ষষ্ঠ ম্যাচে বাংলাদেশ দলের বিপক্ষে	এশিয়া, কাপের, ষষ্ঠ, ম্যাচে, বাংলাদেশ, দলের,বিপক্ষে
রাজধানীতে মাদক বিরোধী বিশেষ অভিযান পরিচালনা	রাজধানীতে, মাদক, বিরোধী, বিশেষ, অভিযান, পরিচালনা

Figure 3.4: Bangla Tokenized Text

Text	Tokenized Text
Pope Francis Just Called Out Donald Trump During His Christmas Speech	Pope, Francis, Just, Called, Out, Donald, Trump, During, His, Christmas, Speech.
10 Ways America Is Preparing for World War 3	10, Ways, America, Is, Preparing, for, World, War, 3.
Coronavirus can be spread through contact with food or food packaging	Coronavirus, can, be, spread, through, contact, with, food, or, food, packaging.

Figure 3.5: English Tokenized Text

অতএব, অথচ, অথবা, অনুযায়ী, অনেক, অনেকে, অনেকেই, অন্তত, অন্য, অবধি, অবশ্য, অর্থাৎ, আই, আগামী, আগে, আগেই, আছে, আজ, আদ্যভাগে, আপনার, আপনি, ই, ইত্যাদি, ইহা, উচিত, উত্তর, এ, এদের, কিন্তু, কী, কে, জানায়, জানিয়ে, জানিয়েছে.....

Figure 3.6: List of Bengali stopwords

3.3.2.2 Stop Word Removal

Information extraction is a result of text data mining where unnecessary words that are very repetitive and common which is actually called the Stop word can create noise. Some of the much generalized examples of stopwords are -a, an, the, on, at, which etc. But there is no universal single list for stop words. These stopwords can take a lot of space and also a large amount of processing time. We can conveniently delete them by saving a list of words.For our English data, we have used the Natural Language Toolkit shortly known as NLTK. In python, we simply needed to improve the nltk and the stopwords for English are already specified in the library. Then from our Headings and Text we can filter out the stopwords.And for removing the bangla stopwords we provided a list of custom stopwords in order to filter out these from the Text and Headlines of the Bengali news. Some of the stopwords that have been filtered out.

3.3.2.3 Punctuation and Garbage character Removal

Cleaning up data means getting rid of the noise in the dataset. In text cleaning, we may have to get rid of all identifiers, HTML instances, punctuation, non-alphabets and all other characters that might not be part of the language. Some of the punctuation that has been considered while filtering them out are shown in fig. 3.7. As we have considered only news articles from different news portals there were no emoticons or characters (which are very much common while working with social media data) found upon manually searching. So, we didn't need to perform any procedure to filter out them. These punctuation and garbage characters can take up a lot of space and increase processing time and as in our model

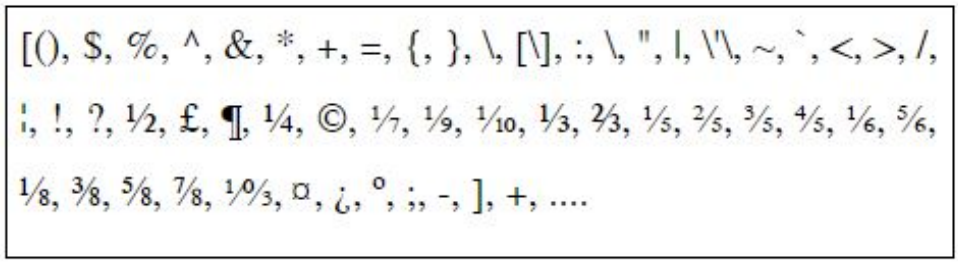

Figure 3.7: Example of Punctuation and garbage characters

we are considering the text of our news we filtered out these garbage characters from our dataset to make this a clearer one.

3.3.2.4 Stemming

Stemming is the act of reducing a word to its word stem which is appended to suffixes and prefixes or to the roots of words such as lemmas. Stemming is important in understanding and processing the natural language. Stemming can be done by defining some rules. In our pre-processing step we have used The Porter Stemming Algorithm. This is a method of deleting the commoner morphological and inflexional endings from English words. The main use of this is a part of a process of standardization of terms usually performed when setting up information recovery systems. This specific part of pre-processing has helped us to have higher accuracies in some of the models.

3.3.3 Feature Extraction

Having a huge amount of data it is very important to get rid of unimportant and redundant information and to prepare the dataset that can be applicable to any other machine learning models or applications or algorithms. Due to the huge amount of text data produced in many ways, such as social networks, hospital reports, news media, the NLP has gained so much attention in recent years. It is impossible for humans to go through all of the text data and locate useful information and coordinate vast volumes of data. To have an overview the method of translation and representation of these data involves measuring word frequencies from the text and in all of the data set. Hence, extracting the correct information is necessary. The mapping of real valued vectors from textual data

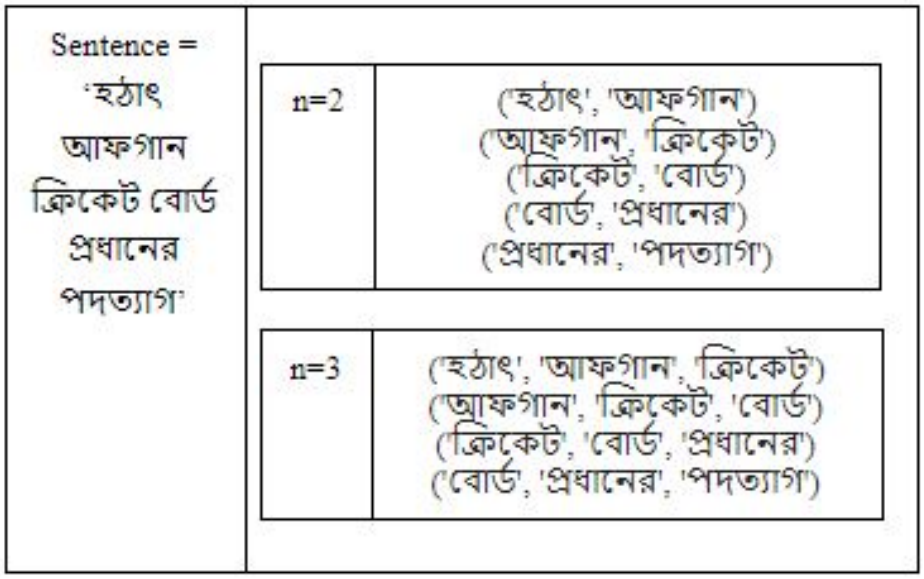

Figure 3.8: Example of N-Gram feature extraction from Bangla

is called feature extraction. The widely used method of extraction of the feature is TF-IDF. TF-IDF technique is improved for extraction of features for greater performance of a model [26, 27]. We have used TF-IDF alngside the N-Gram model.

3.3.3.1 N-Gram

It is merely a sequence of n items from a text corpus or a document. These items can be words, letters, phonemes, syllables or base pairs as needed. This is a very popular approach in the field of Natural Language Processing (NLP). Speech or text corpus is used to get the N-Grams. These have been used in some of the most successful language models. N=1, 2, 3 denotes the unigram, bigram, trigram respectively. A bigram (n=2) is a series of two adjacent components that are usually letters, syllables, or words from a set of tokens. The frequency distribution of each bigram in a string is widely used in many text analysis including computational linguistics, cryptography, voice recognition, etc. A trigram (n=3) is a series of two adjacent components. These are also used in the interpretation of natural language for the mathematical study of texts and for the detection and usage of ciphers and codes in cryptography.

In our work, we have used word-based N-Gram features to see how unigram,

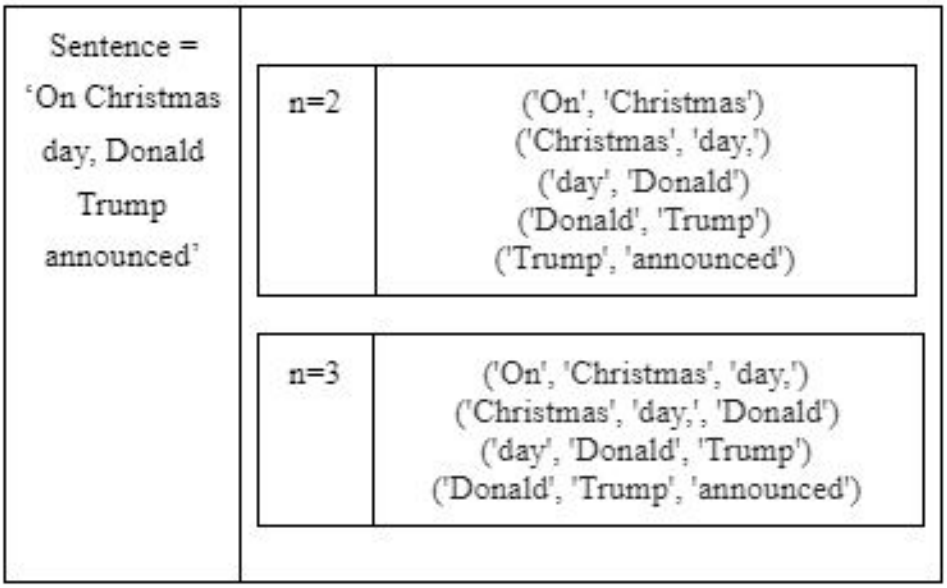

Figure 3.9: Example of N-Gram feature extraction from English

bigram and trigram impacts the overall performance of the model and this was our one of the feature extractors. And, as we have observed while we used bigrams in our six different models we started getting better accuracy with an enriched model but after using the trigrams and four-grams the model performance declined though in a very small range. The N-grams along with the TF-IDF extracted the features for our detection model. Our model is trained with these and later on based on these training experience the model identified fake and authentic news quite successfully. In fig. 3.8 a basic example of N-Gram feature extraction from Bangla sentence and in fig. 3.9 a basic example of N-Gram feature extraction from English sentence is shown.

3.3.3.2 TF-IDF

It stands for Term frequency-Inverse Document Frequency. The value of the TF IDF increases proportionally to the amount of times a term appears in the text and decreases with the number of documents containing the word in the corpus.

```
tf (wd, doc) =fdoc(wd) : Frequency of wd in document doc
idf (wd,DOC)=log1+ |DOC|1+ df (doc,wd)
tfidf (wd,doc, DOC) = tf (wd, doc)  idf (wd, DOC)
```

It consists of two subparts. And they are-1.TF 2. IDF. Actually, in simple words TF-IDF is a product of TF and IDF.

Term Frequency (tf)

Term frequency determines how often a word (wd) appears in the text or document (doc) as a whole. It measures the number of times a word exists in a text document with respect to the overall number of words in the following text document.

Inverse Document Frequency (idf)

The inverse frequency of a text is an indicator of whether a word (wd) in the whole text corpus is unusual or common across the records. The symbol DOC is the representation of the entire text corpus means the number of documents existing there in the text corpus in total. And, "df (doc,wd)"represents the number of documents in which the word appears in. Then as a result of idf (wd, DOC) a value is achieved which is logarithmically scaled and it denotes the number of documents in the corpus divided by the number of times word wd appears in the entire corpus.
A word which is quite prevalent in the entire text corpus will have a lesser TF-IDF value because of the denominator of idf.

We have also applied Bag-of-words which is one of the most common methods in text analysis and specifically in feature extraction, it didnt provide us expected performance as it captures some issues which are not so prominent in the entire dataset. It doesn't care how many times a word appears or the order in which the words appear, all that matters is if the word is in a list.

3.4 Machine Learning Classifiers

The process of predicting the class of given data points is known as classification. Targets, names, and groups are all terms used to describe classes. The role of approximating a mapping function (f) from input variables (X) to discrete output variables is known as classification predictive modeling (y).

In machine learning, a classifier is an algorithm that automatically sorts or categorizes data into one or more classes. One of the most common examples is an email classifier, which scans emails and filters them according to whether they are spam or not. Machine learning algorithms are useful for automating tasks that were previously performed by hand. They will save a lot of time and resources while also increasing the efficiency of companies.

There are several classification algorithms available today, and it is impossible to determine which is superior to the others. It is dependent on the application and the quality of the data set available. For our bilingual fake news detection model, we used six separate machine learning algorithms. They are as follows: i) Logistic Regression ii) Linear Support Vector Classification iii) Decision Tree iv) Random Forest Classifier v) Naive Bayes Multinomial vi) Passive Aggressive.

3.4.1 Logistic Regression

Logistic Regression is a Machine Learning algorithm that is used to solve classification problems. It is a predictive analysis algorithm that is based on the probability principle. It is the method of choice for binary classification issues (problems with two class values).

The logistic function, which is at the heart of the system, is called logistic regression. The logistic function, also known as the sigmoid function, was created by statisticians to explain the properties of population growth in ecology, such as how it rises rapidly and eventually reaches the environment's carrying capacity. It's an S-shaped curve that can map any real-valued number to a value between 0 and 1, but never exactly between those two points.

$$1/(1 + e^{-}value)$$

Where e is the natural logarithms' base (Euler's number or the spreadsheet's EXP() function).

It's important to remember that logistic regression can only be used when the

target variables are discrete, and that if the target value is in a set of continuous values, logistic regression should be avoided. When using logistic regression, a threshold is typically defined, indicating at what value the example will be assigned to one of two classes.

3.4.2 Linear SVC(Support VectorClassifier)

A Linear SVC's (Support Vector Classifier) goal is to suit the data provided and return a "best fit" hyperplane that divides or categorizes the data. After obtaining the hyperplane, some features can be fed into the classifier to determine the "predicted" class. The SVM module (SVC, NuSVC, etc.) wraps the libsvm library and supports several kernel, while LinearSVC is based on liblinear and only supports one kernel. Scikit-learn is a popular library for implementing machine learning algorithms in Python. SVM is also included in the scikit-learn library, and we use it in the same way (Import library, object creation, fitting model and prediction).

3.4.3 Decision Tree Classifier

Decision Tree is a supervised learning method that can be used to solve both classification and regression problems, but it is most often used to solve classification issues. Internal nodes represent dataset attributes, branches represent decision laws, and each leaf node represents the outcome in this tree-structured classifier. The following are two reasons to use the Decision Tree:

- Decision Trees are designed to imitate human thinking abilities when making decisions, making them simple to comprehend.
- Since the decision tree has a tree-like structure, the logic behind it is simple to comprehend.

Algorithm: The algorithm for predicting the class of a given dataset in a decision tree begins at the root node of the tree. This algorithm compares the values of the root attribute with the values of the record (real dataset) attribute and then follows the branch and jumps to the next node based on the comparison. The algorithm compares the attribute value with the other sub-nodes and moves on

to the next node. It repeats the loop until it reaches the tree's leaf node. The following algorithm explains the entire process:

- Step 1: Start with the root node, which contains the entire dataset, says S.
- Step 2: Using the Attribute Selection Measure, find the best attribute in the dataset.
- Step 3: Subdivide the S into subsets that contain the best attribute's possible values.
- Step 4: Create the node of the decision tree that contains the best attribute.
- Step 5: Generate new decision trees in a recursive manner using the subsets of the dataset generated in step 3. Continue this process until the nodes can no longer be classified, at which point the final node is designated as a leaf node.

3.4.4 Random Forest Classifier

Random forests is a method for supervised learning. It has the ability to be used for both classification and regression. It's also the most adaptable and user-friendly algorithm. The trees make up a forest. A forest is said to be more durable the more trees it has. Random forests generate decision trees from randomly chosen data samples, obtain predictions from each tree, and vote on the best solution. It also serves as a good indicator of the value of the function.

Random forests can be used for a number of tasks, including recommendation engines, image recognition, and feature selection. It can be used to detect fraudulent behavior, recognize loyal loan applicants, and forecast diseases. The Boruta algorithm, which selects important features in a dataset, is built around it.

It's theoretically an ensemble method of decision trees generated on a randomly split dataset (based on the divide-and-conquer approach). The forest is the name given to a group of decision tree classifiers. Individual decision trees are created for each attribute using an attribute selection indicator such as knowledge benefit, gain ratio, or Gini index. Each tree is based on a separate random sample. Each tree votes in a classification problem, and the most common class is chosen as

the final result. In the case of regression, the final result is the sum of all the tree outputs. In comparison to other non-linear classification algorithms, it is both simpler and more efficient.

Algorithm: There are four phases to it.

- Take random samples from a collection of results.
- For each sample, create a decision tree and get a prediction result from each decision tree.
- Consider voting for each expected outcome.
- As the final prediction, choose the prediction with the most votes.

Random Forest has a number of advantages.

- Since there are several trees and each tree is trained on a subset of data, the random forest algorithm is not biased. The random forest algorithm, in essence, relies on the "crowd's" power; as a result, the algorithm's overall bias is reduced.
- This algorithm is extremely dependable. Even if a new data point is added to the dataset, the overall algorithm is unaffected and although new data may affect one tree, it is extremely unlikely to affect all trees.
- When you have both categorical and numerical elements, the random forest algorithm works well.
- When data has missing values or has not been scaled well, the random forest algorithm works well (although we have performed feature scaling in this article just for the purpose of demonstration).

And some disadvantages that should be considered as well.

- The complexity of random forests is one of their big drawbacks. Due to the large number of decision trees combined, they needed significantly more computational resources.
- They take much longer to train than other comparable algorithms due to their complexity.

3.4.5 Multinomial Naive Bayes

The Multinomial Naive Bayes algorithm is a probabilistic learning method popular in Natural Language Processing (NLP). The algorithm predicts the tag of a file, such as an email or a newspaper post, using the Bayes theorem. It calculates each tag's probability for a given sample and outputs the tag with the highest probability.

The Naive Bayes classifier is made up of a number of algorithms that all have one thing in common: each feature being categorized is unrelated to any other feature. A feature's presence or absence has no bearing on the presence or absence of another feature.

The Naive Bayes algorithm is a powerful tool for analyzing text data and solving problems with multiple groups. Since the Naive Bayes theorem is based on the Bayes theorem, it is necessary to first understand the Bayes theorem principle. The Bayes theorem, which was developed by Thomas Bayes, measures the probability of an occurrence occurring based on previous knowledge of the event's conditions. It's based on the formula below:

$$P(a|b) = P(a) * P(b|a)/P(b)$$

Here,
P(b) = prior probability of b
P(a) = prior probability of class a
P(b|a) = occurrence of predictor b given class a probability
This formula aids in determining the likelihood of tags in the document.

The following are some of the benefits of the Naive Bayes algorithm:

- It is simple to implement since only the probability should be calculated.
- Both continuous and discrete data can be used for this algorithm.
- It's straightforward and can be used to forecast real-time application.
- It's highly scalable and can handle massive datasets with ease.

The following are some of the drawbacks of the Naive Bayes algorithm:

- This algorithm's prediction accuracy is lower than that of other probability algorithms.
- It isn't appropriate for regression. The Naive Bayes algorithm can only be used to classify textual data and cannot be used to predict numerical values.

3.4.6 Passive Aggressive Classifier

For large-scale learning, passive-aggressive algorithms are commonly used. It's one of the few so-called "online-learning algorithms." In contrast to batch learning, where the entire training dataset is used at once, online machine learning algorithms take the input data in a sequential order and update the machine learning model step by step. This is particularly useful in cases where there is a large volume of data and training the entire dataset is computationally impossible due to the sheer scale of the data. An online-learning algorithm would simply obtain a training example, update the classifier, and then discard the example.

Detecting fake news on a social media platform like Twitter, where new data is introduced every second, is a great example of this. The amount of data needed to dynamically read data from Twitter on a continuous basis would be enormous, so an online-learning algorithm would be perfect. In the sense that they don't need a learning rate, passive-aggressive algorithms are close to Perceptron models. They do, however, have a regularization parameter.

The algorithms are known as passive-aggressive because:
Passive: Hold the model and make no adjustments if the assumption is accurate. i.e., the data in the example is insufficient to trigger any model changes.
Aggressive: Make adjustments to the model if the prediction is incorrect. i.e., a change to the model might fix the issue.

3.5 Implementation

There are some initial requirements while setting up for a fake news detection system. For successful completion of this work these pre-requisites are must. The necessary tools both hardware and software for accomplishing the novel task are listed below:

3.5.1 Hardware Requirements

The minimum and suggested hardware specifications for the bilingual fake news detection system are listed below.

- **Processor and Graphics** - Intel Core i7 + NVIDIA GeForce MX450
- **Display** - 17.3" diagonal FHD, IPS, WLED-backlit, edge-to-edge glass (1920 x 1080)
- **Memory** - 32 GB DDR4-3200 SDRAM (2 x 16 GB)
- **Storage**- 512 GB PCIe NVMe M.2 SSD + Intel Optane memory 32 GB
- **Primary Battery** - 4-cell, 55 Wh Li-ion
- **Keyboard**
- **Precision Touchpad Support / Mouse**

3.5.2 Software Requirements

The minimum and suggested software specifications for the bilingual fake news detection system are listed below.

- **Operating System** - Windows 10 Pro 64
- Python 3.0
- Google Colab
- jupyter notebook
- Sublime text

- Microsoft Excel
- Microsoft Powerpoint

3.5.3 Library requirements

We have used python as our programming language and there are some must have libraries that we have implemented for accomplishing our novel task. They are listed below:

- **Scikit-learn** - It's a machine learning library written in Python. It is, more simply, a collection of easy and effective data mining and data analysis tools. Many other Python libraries, such as matplotlib and plotly for plotting, numpy for array vectorization, pandas dataframes, scipy, and others, work well with Scikit-learn. We can use either pip or conda to install this library on our system. For various types of algorithms, Scikit-Learn provides a model selection option. It comes with pre-built classification, clustering, and other models. Scikit-Learn also includes datasets, preprocessing, feature selection, and other tools.
- **NLTK** - The Natural Language Toolkit is a Python programming language open source library. The Natural Language Toolkit (NLTK) is a Python programming framework for working with human language data in statistical natural language processing (NLP). It includes tokenization, parsing, grouping, stemming, tagging, and semantic reasoning text processing libraries.

3.6 Conclusion

After deciding on our dataset, we pre-processed the text dataset to create our bilingual fake news detection model. Feature extractors are used to extract features from our bilingual text dataset after it has been cleaned. The detection model is trained using these features. The model is then able to conduct test operations on some unlabeled data and distinguish false and authentic news based on its previous experience after completion of training.

Different machine learning models have different merits and demerits and they have their own tricks and contributions to text analysis tasks. That's why we compared six machine learning algorithms on our dataset to evaluate the performance on the data and to check their ability to detect fake and authentic news.

Chapter 4

Results and Discussions

4.1 Introduction

For developing a bilingual fake news detection model we have used six different machine learning algorithms using different tools for pre-processing, cleaning and two different models namely N-grams and TF-IDF for feature extraction. These six models give us different results which are discussed in this following section.

4.2 Dataset Description

We required a dataset that included both English and Bengali texts for our novel model. As a result, we combined two different datasets after extensive processing.

Column name	Number of Contents
Title	42324 non-null object
Text	42324 non-null object
Category	42324 non-null object
Date	42324 non-null object
Label	42324 non-null object

Table 4.1: Bilingual Dataset

We had to remove and rename several columns in order to merge two datasets in two languages. In our final bilingual dataset there are 42324 data in total. The bilingual data frame contained mainly 5 columns in total which we have chosen for training and testing our model. Following that, we worked with the

Title, Text, Category, Date, and Label. The data which were authentic has been labeled with 1 and the data which were fake has been labeled with 0.

4.3 Impact Analysis

The introduction of the World Wide Web and the widespread adoption of social media networks (such as Facebook and Twitter) paved the way for unprecedented levels of information sharing in human history. Consumers are generating and exchanging more content than ever before thanks to social media sites, some of which is deceptive and has little correlation to reality.

A novel technique like ours can impact the web users in a very positive way. An fake news detection system can ensure safe internet surfing for everyone. Many times trauma and panic is generated due to some fake viral news which can harm the mental health of readers greatly. The many losses occurred to the social beings are yet unknown. It can harm the reputation of a renowned person or institution which can be stopped if we futher work with this project and eventually build a system that will automatically identify the veracity of a news.

4.3.1 Social and Environmental Impact

Our ability to make decisions is largely determined by the information we absorb; our worldview is influenced by the information we consume. There is mounting evidence that people have responded belligerently to news that later turned out to be false [28, 29]. The spread of the novel corona virus is a recent example, with false claims about the virus's origin, existence, and actions spreading around the Internet [30]. Fake news like this can lead to social unrest that can become uncontrollable at times. This is when advanced frameworks, such as ours, will step in to prevent any societal or global imbalances.

4.3.2 Ethical Impact

Yellow journalism can be greatly regulated, and more journalists will begin to adhere to the rules and regulations before publishing a story. The responsible

board will strictly follow the ethical guidelines to ensure authenticity. Ethics clearance will be given top priority before any news is published, as the fake news identification system will automatically filter out any fake news in future ventures.

4.4 Evaluation of Framework

We have created a bilingual framework for detecting fake news from online news portals. Our data pre-processing steps include- 1.Tokenization 2.Stemming 3.Stop-words Removal 4.Punctuation Removal. After successful completion of the pre-processing steps we extracted feature score using N-Grams models along with TF-IDF.

English news article	**Label**	**N-Gram Features**	**TF-IDF score**	**Feature score**
'This is an old car. She has a new car'	Fake	['car', 'car she', 'new', 'new car', 'old', 'old car', 'she'.....]	[0.2540002, 0.12700013, 0.12700013..]	1.car 0.254000 2. car she 0.127000. ..
Bengali news article	**Label**	**N-Gram Features**	**TF-IDF score**	**Features score**
'ভাল কাজ করো ভাল ফল পাবে'	Authentic	['ভাল','কাজ', করো', 'করো ফল', 'ফল'...]	[0.2540002, 0.12700013, 0.12700013..]	1.ভাল 0.254000 2. ভাল কাজ 0.127000. ..

Figure 4.1: Example of feature score generation

After extracting the feature scores we have fed them into our model for training purpose.Six different machine learning algorithms trained on these data in their own way of training and later on performed detection operations on test data.

Linear SVC performed best for our data. If we use bigrams it gives the highest

accuracy and if we use trigrams for feature extraction it gives us the highest F1 score. Logistic Regression showed better accuracy with bigrams and it has a great performance too. Decision Tree Classifier has shown highest accuracy with bigrams but the F1 Score doesnt show much of the difference.

Random forest performed very poorly using the data of the Headline column. But, it showed a great F1 score while analyzing Text data. The reason behind this is, while training on the headlines there were less words to consider than the texts of the news. On the other hand, Random Forest Classifier showed pretty good results while training on the headlines too.

The Multinomial Naive Bayes comparatively showed poor results on both the headlines and the texts. The highest accuracy we have got is 87.43% with bigrams that too using the headlines. The Passive Aggressive Classifier showed performance almost as good as Linear SVC both on headings and texts as shown in Table 15. The F1 score denotes the model performs really satisfactory while using trigrams.

4.5 Evaluation of Performance

Various evaluation criteria have been used to measure the efficiency of algorithms for the fake news detection problem. These metrics are widely used in the machine learning community and enable us to assess a classifier's output from various angles.

Accuracy is a metric that compares the resemblance between expected and actual false news. Precision is a metric that calculates the percentage of all detected false news that is annotated as fake news, resolving the critical issue of determining which news is fake. Due to the distorted nature of fake news datasets, a high level of accuracy may be easily accomplished by making less optimistic forecasts. As a result, recall is used to gauge sensitivity, or the percentage of annotated fake news stories that are predicted to be fake news. F1 is used to combine precision and recall, resulting in a prediction value for detecting false news. The higher the value for Precision, Recall, F1, and Accuracy, the better the results.

- **True Positive(TP)**- When the news article which was predicted as fake news is eventually labeled as fake news
- **True Negative(TN)**- When the news article which was predicted as true news is eventually labeled as true news
- **False Negative(FN)**- When the news article which was predicted as true is eventually labeled as fake news
- **False Positive (FP)**- When the news article which was predicted as false is eventually labeled as true news

Accuracy- It is the most common indicator of performance. And this denotes the ratio of correctly predicted observations and total observations. Only higher accuracy doesnt always prove that the model performs really well. But, if the dataset is well balanced then definitely accuracy is the best choice. We have measured the accuracy of each of our models along with the other performance metrics.

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

Precision- It is also called the positively predicted value. Precision is the ratio of accurately forecast positive outcomes to overall expected positive results. A perfect 1.0 accuracy score means any search result was relevant.

$$Precision = \frac{|TP|}{|TP| + |FP|}$$

Recall-It is also known as sensitivity. Recall is the percentage of correctly expected positives to the real class findings. But, Precision is more important than recall you choose to get less False Positives off in exchange to get more False Negatives.

$$Recall = \frac{|TP|}{|TP| + |FN|}$$

F1 Score- In most of the cases, F1 Score is more useful than accuracy when the

dataset is not that balanced. Basically, its the weighted average of precision and recall. It takes both the false positives and false negatives into account which makes it a bit more complicated than accuracy. A F1 score of 1.0 is considered as perfect.

$$F1\ Score = 2.\frac{Precision \cdot Recall}{Precision + Recall}$$

We evaluated the output of our six machine learning models using not only accuracy but also using precision, recall, and F1 score as we don't have equivalent data for Bengali and English, and our dataset is imbalanced based on the labels.

Table 4.2: Experiments results of six different Machine Learning models using the main text of the news articles.

Machine Learning Model	N-Grams	Precision	Recall	F1 Score	Accuracy
Logistic Regression	Unigram	0.92	0.92	0.92	92.23%
	Bigram	0.92	0.92	0.92	92.47%
	Trigram	0.92	0.92	0.92	91.99%
Linear SVC (Support Vector Classifier)	Unigram	0.93	0.93	0.93	92.83%
	Bigram	0.89	0.89	0.89	93.29%
	Trigram	0.93	0.93	0.93	92.83%
Decision Tree Classifier	Unigram	0.92	0.92	0.92	91.53%
	Bigram	0.89	0.89	0.89	91.68%
	Trigram	0.91	0.92	0.92	91.61%
Random Forest Classifier	Unigram	0.91	0.91	0.91	91.27%
	Bigram	0.91	0.91	0.91	90.79%
	Trigram	0.90	0.90	0.90	90.15%
Multinomial Naive Bayes	Unigram	0.84	0.84	0.84	83.52%
	Bigram	0.81	0.81	0.81	81.11%
	Trigram	0.82	0.81	0.80	80.7%
Passive Aggressive Classifier	Unigram	0.92	0.92	0.92	92.38%
	Bigram	0.89	0.89	0.89	93.09%
	Trigram	0.93	0.93	0.93	93.11%

Table 4.2 represents the overall data analyzing the text data of news articles. The model was also trained on the headlines of the news articles but as there are less number of words in the headlines it performed poorly. So, the results of models that have been trained on the text body is represented only. In most of

the cases, the model performed really well where the precision, recall, F1 score is almost perfect sometimes.

4.5.1 Evaluation of performance of Logistic Regression

Probability theory is used to build the predictive analysis algorithm called logistic regression. It is the method of choice when dealing with binary classification problems (problems with two class values). For developing our framework we have trained this model using the headline of the text as well as the main text of the news. The result is discussed below.

Table 4.3: Experiment result of Logistic Regression using the main text and headline of the news articles.

Part of News	**N-Grams**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Main Text	Unigram	0.92	0.92	0.92	92.23%
	Bigram	0.92	0.92	0.92	92.47%
	Trigram	0.92	0.92	0.92	91.99%
Headlines	Unigram	0.88	0.86	0.87	87.6%
	Bigram	0.88	0.87	0.87	87.71%
	Trigram	0.88	0.88	0.88	87.63%

From 4.3 it is easily understandable that our model performs well on main text content of the news as it can train on more words. If the model trains on the headline portion of a news content it will perform comparatively poorly.

4.5.2 Evaluation of performance of Linear SVC(Support VectorClassifier)

Support-vector machines are supervised learning models that analyze data for classification and regression analysis through related learning algorithms.

Table 4.4: Experiment result of Linear Support Vector Classifier using the main text and headline of the news articles.

Part of News	**N-Grams**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Main Text	Unigram	0.93	0.93	0.93	92.83%
	Bigram	0.89	0.89	0.89	93.29%
	Trigram	0.93	0.93	0.93	92.83%
Headlines	Unigram	0.89	0.89	0.89	88.55%
	Bigram	0.89	0.89	0.89	89.05%
	Trigram	0.89	0.89	0.89	88.85%

From 4.4 we can observe better performance using trigram and when the model is trained on the main text of the news.

4.5.3 Evaluation of performance of Decision Tree Classifier

Decision Tree performs very poorly while it is trained on the headline of the news contents with a F1 score of 0.71 only. And, it gives us a great performance using both unigrams and trigrams while training on the main text.

Table 4.5: Experiment result of Decision Tree classifier using the main text and headline of the news articles.

Part of News	**N-Grams**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Main Text	Unigram	0.93	0.93	0.93	92.83%
	Bigram	0.89	0.89	0.89	93.29%
	Trigram	0.93	0.93	0.93	92.83%
Headlines	Unigram	0.84	0.84	0.84	83.84%
	Bigram	0.76	0.76	0.76	76.13%
	Trigram	0.72	0.71	0.71	71.18%

4.5.4 Evaluation of performance of Random Forest Classifier

Just like the previous machine learning models Random Forest outperformed when trained on the main text of a news article. It gives great performance using both unigram and trigram as shown in 4.6

Table 4.6: Experiment result of Random Forest classifier using the main text and headline of the news articles.

Part of News	**N-Grams**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Main Text	Unigram	0.93	0.93	0.93	92.83%
	Bigram	0.89	0.89	0.89	93.29%
	Trigram	0.93	0.93	0.93	92.83%
Headlines	Unigram	0.87	0.87	0.87	87.37%
	Bigram	0.84	0.81	0.82	83.63%
	Trigram	0.82	0.81	0.81	81.44%

4.5.5 Evaluation of performance of Multinomial Naive Bayes

The Multinomial Naive Bayes algorithm is a probabilistic learning method popular in Natural Language Processing (NLP). The algorithm predicts the tag of a file, such as an email or a newspaper post, using the Bayes theorem.

From 4.7 it is observable that this algorithm performed pretty well while identifying fake news from bilingual dataset using trigram and like other models it also performed better with more contents for training purpose.

Table 4.7: Experiment result of Multinomial Naive Bayes classifier using the main text and headline of the news articles.

Part of News	**N-Grams**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Main Text	Unigram	0.93	0.93	0.93	92.83%
	Bigram	0.89	0.89	0.89	93.29%
	Trigram	0.93	0.93	0.93	92.83%
Headlines	Unigram	0.87	0.87	0.87	87.3%
	Bigram	0.88	0.87	0.87	87.43%
	Trigram	0.87	0.87	0.87	86.96%

4.5.6 Evaluation of performance of Passive Aggressive Classifier

One of the incremental learning algorithms available is passive-aggressive classification, which has a closed-form update rule and is very easy to implement. The basic idea is that with each misclassified training sample it receives, the classifier changes its weight vector in an attempt to correct the problem.

From 4.8 It was discovered that, despite performing admirably when trained on the headlines of news stories, it outperformed itself when trained on the main text of the news.

Table 4.8: Experiment result of Passive Aggressive classifier using the main text and headline of the news articles.

Part of News	**N-Grams**	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Main Text	Unigram	0.93	0.93	0.93	92.83%
	Bigram	0.89	0.89	0.89	93.29%
	Trigram	0.93	0.93	0.93	92.83%
Headlines	Unigram	0.87	0.87	0.87	86.87%
	Bigram	0.89	0.89	0.89	88.97%
	Trigram	0.89	0.89	0.89	88.87%

4.6 Conclusion

The classification result for Bilingual Fake News Detection is shown in this chapter. The proposed framework's performance is also addressed. The results show that using TF-IDF in conjunction with the N-Gram model makes the Linear Support Vector a good classifier. All of the classifiers performed comparatively better while they trained on large volume of data like the main contents of the news articles. In the previous part, we discussed our work's conclusion.

Chapter 5

Conclusion

5.1 Conclusion

We developed a model to detect fake news from a bilingual dataset in this paper, with a special emphasis on Bengali and English. In addition, we compared and evaluated the output of six different machine learning techniques using various feature extraction techniques. This model can also be used to create apps that filter out fake news from a news portal, regardless of whether the language is Bengali or English. However, when performing this type of mission, data scarcity is often a major concern..

There are numerous fact-checking websites available these days, most of which focus on Bengali news fact-checking in order to improve the internet, but the majority of them are still carried out manually. If an automated mechanism can identify and filter out fake news, it will have a significant impact on the world of online news portals.

After analyzing the performance of six different machine learning models exploring with the N-Gram feature extractor in our work we have got the highest performance with Linear Support Vector Classifier. If we consider accuracy as the performance metric we acquired the highest accuracy using Bi-Grams and if we take F1-score as our performance metric we have acquired the highest accuracy while using Tri-Grams. As our dataset is not balanced F1-Score will be the best evaluation metric for our model. In future, we will be exploring more features and aspects of fake news with an enriched set of data.

5.2 Future Work

Fake news identification is a relatively new field of research.There are plenty of new scopes and aspects to work into. In a survey [16] authors give direction from a data mining standpoint about relevant research areas, open issues, and possible future works. They have divided their suggestions about future works in four different categories. Based on it our work on bilingual fake news detection system can also be divided into these four.

- Data-oriented: Different aspects of fake news data, such as benchmark data collection, psychological validation of fake news, and early fake news identification, could be given more attention in multiple languages. A low resource language like Bangla should be definitely prioritized in the forthcoming works.
- Feature-oriented: More feature extraction techniques can be explored from our bilingual dataset in future works.
- Model-oriented: More realistic and effective models such as supervised, semi-supervised and unsupervised models, can be created for detecting fake news from our bilingual dataset.
- Application-oriented: Applications to filter out the fake news from potential readers can help them having a safe time on the internet and these can be created based on our proposed framework which can work regardless of the language of the text.

References

[1] V. L. Rubin and N. Conroy, 'Discerning truth from deception: Human judgments and automation efforts,' *First Monday*, vol. 17, no. 3, Mar. 2012, ISSN: 13960466. DOI: 10.5210/fm.v17i3.3933. [Online]. Available: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3933 (cit. on p. 1).

[2] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata and R. Procter, 'Detection and Resolution of Rumours in Social Media: A Survey,' *ACM Computing Surveys*, vol. 51, no. 2, p. 32, Apr. 2017. DOI: 10.1145/3161603. arXiv: 1704.00656. [Online]. Available: http://arxiv.org/abs/1704.00656%20http://dx.doi.org/10.1145/3161603 (cit. on p. 1).

[3] C. Shao, G. L. Ciampaglia, O. Varol, K. C. Yang, A. Flammini and F. Menczer, 'The spread of low-credibility content by social bots,' *Nature Communications*, vol. 9, no. 1, Dec. 2018, ISSN: 20411723. DOI: 10.1038/s41467-018-06930-7. arXiv: 1707.07592. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30459415/ (cit. on p. 1).

[4] M. Granik and V. Mesyura, 'Fake news detection using naive bayes classifier,' *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pp. 900–903, 2017 (cit. on p. 1).

[5] *PolitiFact | 2016 Lie of the Year: Fake news*. [Online]. Available: https://www.politifact.com/article/2016/dec/13/2016-lie-year-fake-news/ (visited on 13th Apr. 2021) (cit. on p. 1).

[6] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts and J. L. Zittrain, 'The science of fake news,' *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018, ISSN: 0036-8075. DOI: 10.1126/science.aao2998. eprint: https://science.sciencemag.org/content/359/6380/1094.full.pdf. [Online]. Available: https://science.sciencemag.org/content/359/6380/1094 (cit. on p. 1).

[7] S. Kogan, T. J. Moskowitz and M. Niessner, 'Fake News: Evidence from Financial Markets,' *SSRN Electronic Journal*, Oct. 2018, ISSN: 1556-5068. DOI: 10.2139/ssrn.3237763. [Online]. Available: https://papers.ssrn.com/abstract=3237763 (cit. on p. 1).

[8] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova and Y. Choi, 'Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking,' in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational

Linguistics, 2017, pp. 2931–2937. DOI: 10.18653/v1/D17-1317. [Online]. Available: http://aclweb.org/anthology/D17-1317 (cit. on p. 4).

[9] Y. Zhu, X. Gao, W. Zhang, S. Liu and Y. Zhang, 'A Bi-Directional LSTM-CNN Model with Attention for Aspect-Level Text Classification,' *Future Internet*, vol. 10, no. 12, p. 116, Nov. 2018, ISSN: 1999-5903. DOI: 10.3390/fi10120116. [Online]. Available: http://www.mdpi.com/1999-5903/10/12/116 (cit. on p. 4).

[10] M. d. P. Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodriguez-García, R. Valencia-García and G. Alor-Hernández, 'Automatic detection of satire in Twitter: A psycholinguistic-based approach,' *Knowledge-Based Systems*, vol. 128, pp. 20–33, Jul. 2017, ISSN: 09507051. DOI: 10.1016/j.knosys.2017.04.009 (cit. on p. 4).

[11] F. Islam, M. M. Alam, S. M. Shahadat Hossain, A. Motaleb, S. Yeasmin, M. Hasan and R. M. Rahman, 'Bengali Fake News Detection,' in *2020 IEEE 10th International Conference on Intelligent Systems, IS 2020 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Aug. 2020, pp. 281–287, ISBN: 9781728154565. DOI: 10.1109/IS48319.2020.9199931 (cit. on p. 4).

[12] M. Aldwairi and A. Alwahedi, 'Detecting fake news in social media networks,' in *Procedia Computer Science*, vol. 141, Elsevier B.V., Jan. 2018, pp. 215–222. DOI: 10.1016/j.procs.2018.10.171 (cit. on p. 7).

[13] S. Aphiwongsophon and P. Chongstitvatana, 'Detecting fake news with machine learning method,' *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 528–531, 2018 (cit. on p. 7).

[14] N. J. Conroy, V. L. 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*, ser. ASIST '15, St. Louis, Missouri: American Society for Information Science, 2015, ISBN: 087715547X (cit. on p. 7).

[15] M. Granik and V. Mesyura, 'Fake news detection using naive Bayes classifier,' in *2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 900–903, ISBN: 9781509030064. DOI: 10.1109/UKRCON.2017.8100379 (cit. on p. 7).

[16] K. Shu, A. Sliva, S. Wang, J. Tang and H. Liu, 'Fake news detection on social media: A data mining perspective,' *SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, Sep. 2017, ISSN: 1931-0145. DOI: 10.1145/3137597.3137600. [Online]. Available: https://doi.org/10.1145/3137597.3137600 (cit. on pp. 7, 44).

[17] A. Chakraborty, B. Paranjape, S. Kakarla and N. Ganguly, 'Stop Clickbait: Detecting and preventing clickbaits in online news media,' in *Proceedings*

of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, Institute of Electrical and Electronics Engineers Inc., Nov. 2016, pp. 9–16, ISBN: 9781509028467. DOI: 10.1109/ASONAM.2016.7752207. arXiv: 1610.09786 (cit. on p. 8).

[18] K. Popat, S. Mukherjee, J. Strötgen and G. Weikum, 'Credibility assessment of textual claims on the web,' in *International Conference on Information and Knowledge Management, Proceedings*, vol. 24-28-October-2016, New York, NY, USA: Association for Computing Machinery, Oct. 2016, pp. 2173–2178, ISBN: 9781450340731. DOI: 10.1145/2983323.2983661. [Online]. Available: https://dl.acm.org/doi/10.1145/2983323.2983661 (cit. on p. 8).

[19] S. Shojaee, M. A. A. Murad, A. B. Azman, N. M. Sharef and S. Nadali, 'Detecting deceptive reviews using lexical and syntactic features,' in *International Conference on Intelligent Systems Design and Applications, ISDA*, IEEE Computer Society, Oct. 2014, pp. 53–58, ISBN: 9781479935161. DOI: 10.1109/ISDA.2013.6920707 (cit. on p. 8).

[20] H. Ahmed, I. Traore and S. Saad, 'Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques,' in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10618 LNCS, Springer Verlag, 2017, pp. 127–138, ISBN: 9783319691541. DOI: 10.1007/978-3-319-69155-8_9 (cit. on pp. 8, 14).

[21] W. Y. Wang, '"liar, liar pants on fire": A new benchmark dataset for fake news detection,' in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 422–426. DOI: 10.18653/v1/P17-2067. [Online]. Available: https://www.aclweb.org/anthology/P17-2067 (cit. on p. 8).

[22] N. F. Baarir and A. Djeffal, 'Fake news detection using machine learning,' *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, pp. 125–130, 2021 (cit. on p. 8).

[23] H. E. Wynne and Z. Z. Wint, 'Content based fake news detection using n-gram models,' *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, 2019 (cit. on pp. 8, 9).

[24] I. Ahmad, M. Yousaf, S. Yousaf and M. O. Ahmad, 'Fake News Detection Using Machine Learning Ensemble Methods,' *Complexity*, vol. 2020, 2020, ISSN: 10990526. DOI: 10.1155/2020/8885861 (cit. on p. 9).

[25] M. Z. Hossain, M. A. Rahman, M. S. Islam and S. Kar, 'BanFakeNews: A dataset for detecting fake news in Bangla,' English, in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 2862–2871, ISBN:

979-10-95546-34-4. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.349 (cit. on pp. 9, 14).

[26] L. H. Patil and M. Atique, 'A novel approach for feature selection method tf-idf in document clustering,' in *2013 3rd IEEE International Advance Computing Conference (IACC)*, 2013, pp. 858–862. DOI: 10.1109/IAdCC.2013.6514339 (cit. on p. 19).

[27] L. Jing, H. Huang and H.-b. Shi, 'Improved feature selection approach tfidf in text mining,' *Proceedings. International Conference on Machine Learning and Cybernetics*, vol. 2, 944–946 vol.2, 2002 (cit. on p. 19).

[28] A. Robb, 'Anatomy of a fake news scandal,' *Rolling Stone*, vol. 1301, pp. 28–33, 2017 (cit. on p. 32).

[29] J. Soll, 'The long and brutal history of fake news,' *Politico Magazine*, vol. 18, no. 12, p. 2016, 2016 (cit. on p. 32).

[30] J. Hua and R. Shaw, 'Corona virus (covid-19) infodemic and emerging issues through a data lens: The case of china,' *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, 2020, ISSN: 1660-4601. DOI: 10.3390/ijerph17072309. [Online]. Available: https://www.mdpi.com/1660-4601/17/7/2309 (cit. on p. 32).